SZAKMAI BESZÁMOLÓ:

1. Bevezetés

Projektünk célja volt a finnugor nyelvészet és a számítógépes nyelvészet egyik ágának, a természetes nyelvek feldolgozásának összekapcsolásával szolgálni a veszélyeztetett oroszországi finnugor nyelvek közül kettő, a manysi és az udmurt támogatását nyelvtechnológiai eszközök létrehozásával. Mindkettő veszélyeztetett nyelv: az UNESCO által szakmai igényességgel létrehozott "Atlasz a világ veszélyeztetett nyelveiről" (UNESCO *Atlas of the World's Languages in Danger,* http://www.unesco.org/languages-atlas/en/atlasmap.html) besorolása szerint az udmurt "mindenképpen veszélyeztetett" ("definitely endangered"), a manysi pedig "súlyosan veszélyeztetett" ("severely endangered").

A természetes nyelvek feldolgozása (Natural Language Processing, NLP) a nyelvészet és az informatika metszéspontjánál elhelyezkedő tudományág, feladata az emberi nyelvek írásbeli és szóbeli formáinak automatikus feldolgozása és generálása. A tudományág egyik fő célkitűzése az ember és ember vagy ember és gép közti kommunikáció hatékonyságának elősegítése, valamint az újszerű technológiákkal és szolgáltatásokkal, mint például helyesírás-ellenőrzőkkel, diktálórendszerekkel kapcsolatos emberi munka megkönnyítése. További cél a különféle fogyatékosságokkal, például látáskárosodással, halláskárosodással, nyelvi kapacitást csökkentő agyi sérüléssel élő személyek segítése.

A természetes nyelvek feldolgozásának technikái és eszközei szintén fontos szerepet játszanak a digitális bölcsészettudományok támogatásában, így például a veszélyeztetett nyelvek megőrzése és támogatása meghatározó tendenciává vált, és népszerűségnek örvend az NLP kutatói közösségében. Az UNESCO jelentése (UNESCO 2003) szerint a jelenleg ismert nyelvek, köztük a veszélyeztetett nyelvek 50-90%-a, valószínűleg el fog tűnni a 21. század végére, így egy veszélyeztetett nyelv bármilyen korábban ismeretlen módon történő kutatása önmagában is indokolt: a nyelvi dokumentáció fontossága mellett a szociolingvisztikai és antropológiai nyelvi kutatások szükségessége, illetve nyelvtechnológiai eszközök készítése a nyelvhasználók és nyelvtanulók számára egyaránt megkérdőjelezhetetlen.

Projektünk fő célja volt nyelvtechnológiai eszközök létrehozása az udmurt és manysi nyelvekre, hogy beszélőik modern, 21. századi módokon tudják használni e veszélyeztetett nyelveket. Különösen fontos ez a fiatal nemzedék képviselői számára: ők egyrészt a "digitális bennszülöttek" populációjába tartoznak, és ezért mindennapos közegük a digitális eszközök használata, másrészt nyelvhasználat szempontjából kiemelt fontosságú az ő udmurt, illetve manysi nyelvhasználatuk támogatása, hiszen ők a nyelvi közösségükben végbemenő nyelvcserefolyamat eredményeképpen tipikusan orosz dominánsak, és közösségük eredeti nyelvét egyre szűkülő nyelvhasználati színtereken használják, s így nyelvtudásuk szintje is alacsonyabb az idősebb generációkénál. Az ő esetükben tehát a nyelvtechnológiai eszközök nem csak a nyelv használatában, de annak visszatanulásában is kulcsszerepet játszhatnak.

A szibériai őshonos népek körében lezajló urbanizációs folyamatokat ritkán tanulmányozták nyelvészeti, néprajzi vagy kulturális antropológiai szempontból, dacára annak, hogy a városi lakosság mellőzésével a kutatható csoportok nagy részétől tekintenek így el. Különösen igaz ez az oroszországi kisebbségek esetében: az obi-ugorok többsége például ilyen módon nem tárgya a nyelvészeti és néprajzi kutatásoknak (Nagy 2016). Ez általában a veszélyeztetett, különösen pedig az obi-ugor nyelvek szociolingvisztikai vizsgálatát, különösen a városi nyelvhasználat és a nyelvi revitalizáció kutatását kiemelten fontossá teszik.

Projektünkben célunk volt még az udmurt és a manysi nyelv beszélőinek szociolingvisztikai vizsgálata, elsősorban digitális nyelvhasználatuk (azaz digitális eszközök által közvetített nyelvhasználatuk) felmérése az általunk létrehozott nyelvtechnológiai eszközök üzembehelyezése előtt és után, hogy visszajelzést kapjunk azok használatáról és hasznosságáról. A szociolingvisztikai felmérések projektünk meg nem valósíthatónak bizonyult komponensei lettek azonban, mivel – számunkra is váratlan mértékben – a beszélők nem mutattak hajlandóságot kérdéseink megválaszolására (erről bővebben alább).

2. Eredményeink: létrehozott nyelvtechnológiai eszközök

2.1. Udmurt szótár

A projektben eredetileg az addig csak nyomtatott formában létező, Kozmács István által írt Udmurt–magyar szótár (2002) digitális változatának elkészítése szerepelt, aminek fejlesztése közben azonban egy udmurt nyelvi korpusz építését is megkezdtük. Finn partnerünk morfológiai elemzővel rendelkező, korpusz alapú nyelvoktató és nyelvfenntartó programon dolgozott. Ennek megfelelően a következő műveletek elvégzésével kezdődött meg a szótár fejlesztése:

(i) Elkészült a nyomtatott szótár eredeti Word kézirata alapján egy olyan fájl, amely alapján a számítógépes fejlesztők szerint, az általuk megadott szerkesztési elveket követve létrejön az udmurt–magyar–udmurt szótár digitálisan használható változata, illetve ebből elkészülhet egy magyar–udmurt nyomtatott szótár is.

(ii) Elkezdődött a korpuszépítés, amelynek alapja Cseke Adrienn egykori szegedi diákunk által összeállított szövegállomány volt, ami az Udmurt Dunnye című napilapnak 2013 szeptemberéig az interneten elérhető minden szövegét tartalmazza.

(iii) Finn partnerünk által fejlesztett elemző tesztelését és validálását is mi végeztük, illetve a korpuszunk számára szolgáltattunk szövegeket.

Miközben ezeket a munkálatokat (i)–(iii) folytattuk, nyilvánosságot kapott a Национальный исследовательский университет «Высшая школа экономики» (Moszkva) keretében működő Школа лингистики munkatársai által fejlesztett udmurt korpusz (http://web-corpora.net/UdmurtCorpus/search/), amely 7,3 millió tokent tartalmaz, s amelyhez egy morfológiai elemző is kapcsolódik. Ez a korpusz tartalmazza a mi rendelkezésünkre álló szöveganyagot, továbbá blogszövegeket és szépirodalmi szövegeket is. Néhány éve emellett a korpusz mellett elérhető egy 65 ezer tokenes beszermán korpusz is (http://beserman.ru/corpus/search/?interface_language=ru), amely mellett korpuszalapú szótár is található. A korpusz fejlesztőivel felvettük a kapcsolatot, Szegeden egy konferencián személyesen is találkoztunk velük, s bár együttműködés ténylegesen nem jött létre a számítógépes fejlesztés területén, a korpuszépítésről letettünk és a saját szótári munkánkhoz az említett korpuszokat használtuk.

A szótár fejlesztése során együttműködést alakítottunk ki Jack Rueterrel (Helsinki Egyetem), akinek a projektje keretében a Giellatekno felületén többnyelvű udmurt szótár munkálatai folytak (ld. http://kyv.oahpa.no/udm/hun/; http://kyv.oahpa.no/hun/udm/). Ezen a felületen az udmurt mellett mari, számi, mordvin, nyenyec és kis balti-finn nyelvek szótárai készülnek. Az udmurt nyelv finn, komi és észt szótára egészül ki a magyar változattal. Rueter rendelkezésünkre bocsátotta azt az udmurt szavakat tartalmazó fájlt, amelyben az általa

megadott szerkesztési elveknek megfelelően kerülnek be a magyar megfelelők. Ez a szólista 51 178 egységet tartalmaz.

A helyzet elemzéséhez még hozzátartozik, hogy az udmurtok körében is folytak/folynak fejlesztések. Ennek eredményeképpen ma már két különböző formában elérhető az udmurt–orosz szótár a világhálón. Az egyik változat egy gyakorlatilag internetes felületen megjelenő nyomtatott szótár, kereshető felületen. A másik egy olyan változat, ahol lehetőség van a keresendő lexikai egység beírására, majd a program a szótár teljes anyagából bemutatja azokat a tételeket, amelyekben az adott szó (nemcsak címszóként) előfordul. A világhálón ugyanitt szabadon elérhetőek az egyes tudományterületek számára készült szakszótárak is, amelyeket általában kereshető szövegként rögzítettek pdf formátumban.

A szótár – később bemutatandó formájában – alapjául szolgálhat egy olyan elemzőnek, amely már nemcsak morfológiai elemző, hanem képes a mondatok elemeit szemantikai szempontból is értelmezni, és így adekvát(abb) elemzést nyújtani. Az elemzők – akár a korábban az MTA-n, akár a jelenleg a finn partnerek által jelenleg készülőben levő, akár az orosz intézményben létrejött -– nem képesek megküzdeni azzal a problémával, amit az udmurt nyelvben rendkívül gyakori szóalakszinonimitás és a produktív képzések okoznak. Az orosz elemző például a *вуттиськем* szóalaknak hét elemzését adja (ennek oka, hogy ez a szóalak lehet a *vutty-*, illetve az ebből képzett *vuttisky-* ige múlt idejű melléknévi igenévi alakja, illetve ezen igéknek több finiti igei alakja is), míg a *вутэм* szólakot csak igei származéknak elemzi, annak ellenére, hogy a korpuszban előforduló hét esetből ötször e szóalak nem a *вуыны* ’érkezik, megérkezik, megjön, befut, beköszönt; megérik, beérik; idekerül’ igéből képzett kauzatív igei participiuma, hanem a *ву* ’víz’ szóból’ származó ’víztelen’ jelentésű melléknév.

Az elemzők többek között azért nem képesek ezt a problémát kezelni, mert a rendelkezésünkre álló szótárak számos esetben a lexikai egységek jelentéseit nem adják meg. Az udmurt–orosz–udmurt szótár gyakorlata, hogy az egy igei tőből produktív képzéssel létrehozható szavakat felsorolja ugyan, de jelentésüket nem adja meg (az udmurtban ugyanis három igeképző produktív módon gyakorlatilag minden igetőből létrehozhat új igét, akár egymással kombinálódva is). A Kozmács István által készített nyomtatott udmurt–magyar szótár hasonlóan jár el. Mivel annak elsődleges célközönsége eredetileg az udmurtul tanuló magyar diákok voltak, ezért a szótár még a képezhető szóalakokat sem tartalmazza, csupán azt az utasítást adja, hogy a képzett alakokat keresse a felhasználó az alapigénél, és a grammatikában található képzőállomány felhasználásával hozza létre a kívánt jelentést. Ez nem magyar felhasználó számára komoly akadályt jelent, ugyanis mivel nem szerepelnek a képzett igék, így nem szerepelnek azok magyar jelentései sem, így a szótár nem tartalmazza például a ’csináltat’ jelentésű lexémát sem udmurtul, sem magyarul. Az udmurt–magyar szótár nem tartalmazza az udmurt nyelvben meglévő orosz jövevényszavak igen nagy számban létező elemeit sem.

A projekt során megvalósult szótár pótolja ezeket a hiányosságokat. A megvalósítandó feladat egy olyan magyar–udmurt változat előállítása volt, amely megfelel a modern digitális szótár, illetve magának a szótárnak a kritériumainak, azaz korpuszalapon készült és alkalmas valamely morfológiai elemző alapjául szolgálni, illetve amelynek alapján elkészíthető egy olyan elemző, amely már a szemantikai viszonyokat is kezelni tudja. További jelentősége, hogy mivel a szólista, amely alapján készül, gyakorlatilag tartalmazza

az udmurt nyelv eddig feljegyzett minden lexémáját (alapja ugyanis az az udmurt–orosz szótár, amelyben a szerkesztők a korábbi szótári adatok minden elemét egybegyűjtötték, továbbá tartalmaz más forrásokból is elemeket, amelyeket a meglévő szótárakba nem jegyeztek fel), illetve korpusz alapú, ezért olyan módon adja meg a jelentéseket, ahogy az korábban nem történt meg, azaz egy nem hagyományos kétnyelvű szótár készülhetett el. Egy-egy lexikai egység, lexéma esetében a szótár tartalmazza annak szófaját, ragozás és képzés során használatos tövét, kifejezésekben való előfordulásait, ezekkel bemutatva vonzatait, stiláris, használati jellemzőit.

A szótár jelen alakjában, a Rueter-féle felületen még nem rendelkezik az összes felsorolt tulajdonsággal. Ennek oka, hogy a Rueter-féle szerkesztési elvek alapvetően egy szólista jellegű szójegyzéket eredményeznek, miközben esetünkben egy korpuszalapú, korpuszból származó adatokat is tartalmazó szótár elkészítése a cél. A szólista jelleg azt jelenti, hogy az egyes címszavakhoz tartozó, hagyományosan a címszót követő szerkezeteket, összetételeket a szólista az ábécének megfelelően és önállóan tartalmazza, így a szerkesztés meglehetősen időigényes. A kereső felületen a keresett szó beírása során a program felajánl lehetőségeket, így összekapcsolva a listában egymástól távol álló elemeket. További probléma, hogy jelenleg még nem tisztázódott, hogy hozható létre a rendelkezésre álló digitális tartalmakból a tervezett szótárnak akár a digitális felülete, akár a papíralapú változata.

A szótár jelenlegi készültségi állapota: ~20 000 tétel a magyar megfelelőkkel, szófaji besorolással és tövekkel a Rueter-féle anyagban, ami tartalmazza az udmurt–magyar szótár lexikai anyagát. A fejlesztő, Jack Rueter közlése szerint ez az anyag 2018. március 31-től lesz mindenki számára elérhető.

.

2.2. A projekt során elkészített manysi produktumok

2.2.1. A manysi nyelv

A manysi (korábbi elnevezése: vogul) nyelvet Nyugat-Szibériában beszélik, a hanti nyelvvel együtt alkotja az uráli nyelvcsalád obi-ugor ágát. A 2010-es oroszországi népszámlálás eredményei alapján 12 262 ember vallotta magát manysinak, míg a nyelvet mindössze 938 ember beszéli. Összehasonlítva ezeket a számokat a 2002-es népszámlálás adataival -- 11 432 manysi és 2746 beszélő --, megállapítható a korábban is jellemző tendenciák folytatódása: míg a beszélők száma határozottan csökken, addig a népcsoport mérete lassan, de folyamatosan nő. A manysik többsége (10 977 fő) a Hanti-Manysi Autonóm Körzet -- Jugra területén él, fennmaradó hányaduk nagyrészt Oroszország szomszédos közigazgatási egységeiben. A manysi nyelv nyelvjárásokra osztása a 19. század második felében, a kategorizációt lehetővé tévő nyelvi anyag összegyűjtése után vette kezdetét.

A nyugati és oroszországi irodalomban egyaránt leggyakrabban használt felosztás négy nyelvjáráscsoportot különböztet meg a manysiban. Az északi nyelvjáráscsoportba tartozó nyelvváltozatokat a Szoszva és a Ljapin folyók mentén, illetve a Berjozovói körzetben beszélik. Ez a nyelvjáráscsoport (különösen a szoszvai nyelvváltozat) szolgált az irodalmi manysi nyelv alapjául. Az északi nyelvjárásokat erős orosz, valamint komi és nyenyec hatás jellemzi, emellett az északi hanti dialektussal is kapcsolatban álltak.

A nyugati nyelvjáráscsoport a valaha a Lozva folyó középső és alsó folyásánál beszélt nyelvváltozatokból állt, orosz és komi kölcsönhatás jellemezte. A keleti nyelvjárásokat a Konda és a Jukonda folyók mentén beszélték, tatár nyelvi hatások figyelhetők meg rajtuk. A déli nyelvjáráscsoportba tartozó nyelvváltozatokat a Tavda folyó mentén beszélték, itt figyelhető meg a legerősebb tatár nyelvi kölcsönhatás. A déli nyelvjárások a 20. század első felében, a nyugati nyelvváltozatok a 20. század második felében, valószínűleg a '60-as, '70-es években haltak ki, a keleti dialektusnak a kortárs terepmunkaadatok alapján még él néhány beszélője.

2.2.2. A projekt során elkészített nyelvtechnológiai eszközök és adatbázisok

2.2.2.1. Északi manysi korpusz

Statisztikai gépi tanulási módszerrel működő eszközök készítése során az annotált adatbázisok kiemelkedő fontossággal bírnak a természetes nyelvek feldolgozása területén. Munkálataink során ezért létrehoztunk egy az egyetlen manysi nyelvű sajtótermék, a Luima Seripos[1] nevű, kéthetente megjelenő folyóirat szövegei felhasználásával épített északi manysi korpuszt.

Mivel a Luima Seripos újság a kortárs manysi szövegek egyetlen stabil forrása, rendszeresen és kiszámíthatóan elérhető manysi fórumként nemcsak a nyelv leírásának szempontjából kiemelkedő jelentőségű, de hatással van a nyelvhasználatra és a nyelvelsajátításra is. Ennek köszönhetően, vagyis részben azért, mert a Luima Seripos képezi korpuszunk alapját, részben azért, mert a Luima Seriposban használt helyesírás egyre szélesebb körben, például a közösségi médiában is elterjedt, munkánk során a Luima Seripos ortográfiai rendszerét követjük.

A 20. század első évtizedeiig a manysi nyelvről nyelvészek, tudományos kutatók és misszionáriusok készítettek feljegyzéseket, néhány marginális kivételtől eltekintve különböző latin betűs átírásokat alkalmazva. A manysi írásbeliség létrehozásakor az 1930-as években eleinte tovább folytatódott a latin betűs lejegyzés használata, míg a Szovjetunió nyelvpolitikájának változása következtében 1937-től kötelezővé nem vált a cirill alapú ortográfia. 1937 óta a sztenderd manysi helyesírás több alkalommal változott, egyre kidolgozottabbá vált, az orosz nyelvben nem fellelhető fonémák jelölésére használt speciális karakterekkel és diakritikus jelekkel bővült. A jelenleg használt kétféle ortográfia közül a tudományos kiadványok többségében használt írásmód az 1980-as évek végére alakult ki, a közelmúltig minden iskolai tankönyvben is ez szerepelt, míg az ettől csak minimálisan eltérő, a Luima Seripos szerkesztősége és a revitalizációs kísérletekben részt vevő városi aktivisták által használatos, így az általuk írott tankönyvekben bevezetett és az orosz anyanyelvű manysi nyelvhasználók számára ideálisabbnak vélt írásmód rövidebb hagyományra tekinthet vissza.

A korpusz jelenleg a Luima Seripos 1989 óta megjelenő folyóirat 2013 óta kiadott számainak szövegét tartalmazza. A korpusz mintegy 800 000 tokenből áll, XML formátumba konvertálva. A mellékjeles cirill betűket, vagyis a hosszú magánhangzókat két karakterre bontottuk: a tulajdonképpeni magánhangzóra és a hosszúságát jelölő írásjelre, ilyen módon

[1] http://www.khanty-yasang.ru/luima-seripos/archive

lehetővé téve a megfelelő Unicode karakterkódolást (szemben a honlapon található eredeti formátummal, mely egyes gépeken és böngészőkben hibásan jelenik meg).

A cikkek szövege mellett a korpusz a következő metaadatokat is tartalmazza: az újság száma, a megjelenés dátuma, a szerző neve, a cikk címe, a cikkhez vezető link, illetve egy minden cikkhez hozzárendelt egyedi azonosító szám. Az XML fájl a tulajdonnevek, illetve a beágyazott orosz nyelvű szövegbetétek jelölésére külön taget alkalmaz.

A korpuszból egy kisebb minta megtalálható a projekt honlapján itt: http://www.ieas-szeged.hu/finugrevita/results_hu.html

2.2.2.2. Északi manysi szótár

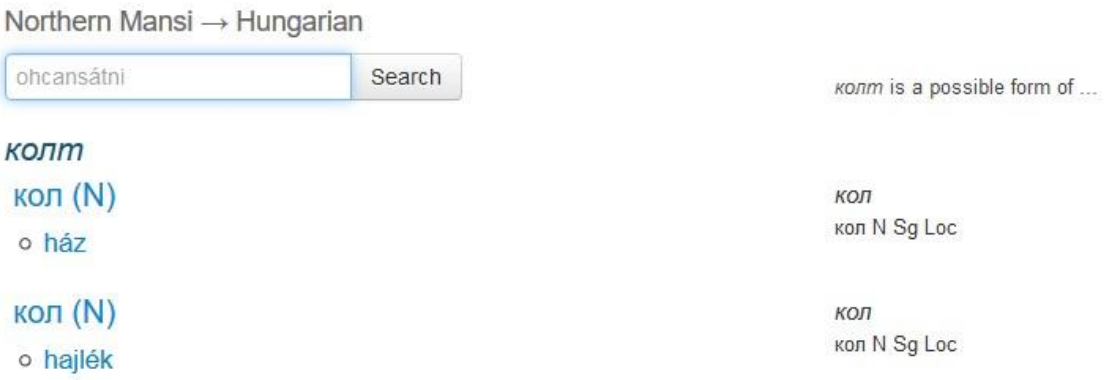
Manysi nyelvű szótárból alig néhány létezik. Az első szisztematikus kutatók, Munkácsi és Kannisto gyűjtése alapján összeállított szótárak (Munkácsi–Kálmán (1986) és Kannisto (2013) ma is fontos eszközei és jelentős adatforrásai a kutatóknak. Mindazonáltal ezek a szótárak nem megfelelőek a kutatók és a manysi nyelvhasználók számára egyaránt hasznos nyelvtechnológiai eszközök létrehozása során, részben azért, mert mindkét szótár különösen aprólékos, mellékjeles latin átírást használ, részben azért, mert mindkét szótár tartalmazza a Munkácsi és a Kannisto terepmunkái óta kihalt nyelvjárások és nyelvváltozatok szóanyagát is, míg hiányoznak belőlük a modern élethez kapcsolódó és annak során használt neologizmusok.

Így munkánk során az európai kutatók által összeállított nagyobb szótárak helyett a manysi kutatók által készített manysi–orosz–manysi (Rombangyejeva–Kuzakova 1982) és orosz–manysi (Rombangyejeva 2005) szótárak felhasználása mellett döntöttünk. A 4000, illetve 11 000 szócikket tartalmazó szótárak mellékjeles cirill írásmódot használnak, tartalmazzák a jelenleg használt kifejezéseket és a manysi közoktatásban és a nyelvvel kapcsolatos szakmai munkák során is általánosan használatosak.

A digitális szótárépítés során a szótárakat előbb automatikus optikai karakterfelismeréssel alakítottuk át, majd a bejegyzéseket kézzel javítottuk, és hozzáadtuk a szóelemek fordítását, ilyen módon kereshető, digitalizált szótárt (Thieberger–Berez 2012) hoztunk létre.

Az online manysi szótár összesen mintegy 15 000 bejegyzést tartalmaz. A két szótárból OCR segítségével kinyert manysi alakokat összevontuk, feltüntetve, mely alak melyik szótárban található meg. A több szinonimát tartalmazó fordításokat különálló elemekként vettük fel. A szótár tartalmazza a manysi szavak szófaját, morfológiai paradigmák szerinti besorolását, és vonzatszerkezetét is. A manysi szóelemeket a már a szótárakban rendelkezésre álló orosz, illetve a projekt során elkészített magyar fordítás egészíti ki. A szótár manysi–magyar verziója a következő linken érhető el: http://vada.oahpa.no/mns/hun/

A szótár keresőjében meg lehet adni szótöveket is, például *кол* ‘ház, hajlék’. Ragozott alakokat is meg lehet adni, például *колт*. Ekkor a program megadja a szótári alakot *кол*, a szófajt zárójelben (N, azaz főnév), illetve hogy a beírt alak milyen számú és személyű és milyen esetraggal áll, például a mi esetünkben ez lehet egy N Sg Loc, azaz egyes számú locativusi esetű főnév.



A program szóalak-generáló funkcióval is rendelkezik. Ha a *кол* (N) feliratra kattintunk, akkor az online felület megadja a főnév összes létező alakját azok morfológiai annotációival együtt.

A szótár manysi–orosz verzióját itt találhatjuk: http://vada.oahpa.no/. Ez a verzió természetesen az elsődlegesen fontos a manysi etnikumú, de a nyelvet nem beszélő olyan egyéneknek, akik elektronikusan könnyen elérhető módon hozzáférnek a szótárhoz, hogy abban (például manysi szöveg olvasásakor) megnézhessenek szavakat, szóalakokat.

2.2.2.3. Északi manysi helyesírás-ellenőrző

Az online északi manysi helyesírás-ellenőrző a Giellatekno norvég fejlesztőivel együttműködésben valósult meg, a következő linken érhető el: http://divvun.org/proofing/online-speller.html A nyelv egy felső legördülő menüben adható meg. A helyesírás-ellenőrző le is tölthető a következő linkről:
http://divvun.org/proofing/proofing.html

Mivel projektünk célja nyílt forráskódú alkalmazások fejlesztése volt, ezért a letölthető helyesírás-ellenőrző is a szintén nyílt forráskódú LibreOffice programhoz telepíthető.

2.2.2.4. Északi manysi elemző

A manysi gazdag morfológiával rendelkező nyelv, tehát az erre a nyelvre készült NLP-eszközök jó morfológiai elemzőt igényelnek. Bár a manysi nyelv folyamatosan és egyre erősödően veszélyeztetett, néhány morfológiai elemző készült már hozzá. Az északi manysihoz ugyan létezett már projektünk előtt is morfológiai elemző, de az latin karaktereket használ és csak a Kálmán-féle *Chrestomathia Vogulica* és *Wogulische Texte mit einem Glossar* szövegeire, illetve Munkácsi szövegeire van optimalizálva. Ezen felül sajnos nem is nyílt forráskódú. Ezért tartottuk fontosnak létrehozni az északi manysira létrehozott morfológiai elemzőnket, amely elérhető a Giellatekno internetes felületéről, ahol a kezelőfelület három nyelven (oroszul, angolul, és finnül elérhető):
http://giellatekno.uit.no/cgi/d-mns.eng.html
http://giellatekno.uit.no/cgi/index.mns.rus.html
http://giellatekno.uit.no/cgi/index.mns.fin.html

A jelenleg rendelkezésre álló véges állapotú eszközök közül a HFST szabványt választottuk, hogy az elemző illeszkedhessen a Giellatekno weboldalán alkalmazott keretbe. Ezt a választást tehát főleg az a törekvés motiválta, hogy ily módon a morfológiai elemző integrálható legyen egy olyan nagyobb rendszerbe, mely egy közös felületen több kisebbségi finnugor nyelvet is kezel.

A morfológiai elemzőben lévő fájlok két kategóriába sorolhatók: tövek és toldalékok. A manysi szavak (szótövek), a morfológiai információkkal és a fordításokkal együtt a szótárban találhatók, valamint a morfológiai szabályok is, amelyek az egyes szótövek különböző inflexiós formáinak elemzéséhez és előállításához szükségesek. A morfológiai szabályok használata érdekében a szótár manysi lexikai elemeit morfológiai kategóriákba soroltuk aszerint, hogy melyik ragozási paradigmához tartoznak. A kategóriák felállítása során a rendelkezésre álló manysi nyelvtanokra (Riese 2001, Rombangyejeva 1973) támaszkodtunk.

A manysi szótövek inflexiós paradigmák szerinti csoportosítását a szavak fonológiai struktúrája alapján végeztük. Ezután minden egyes típushoz inflexiós paradigmákat hoztunk létre. Jelenleg a rendszer 36 névleges és 27 verbális paradigmát tartalmaz. A morfológiai elemző a ragozható szófajok mellett a nem ragozható szófajokat is tartalmazza.

Az elemzőt és szó- és paradigmagenerátort fiatal, orosz-domináns manysik (vagy más nyelvi hátterű kutatók stb.) írott nyelvi szövegek olvasásakor tudják használni: pl. a ragozott alakok elemzésére, lefordítására. Tehát manysi tudásukat gyarapíthatják vele (akár előző manysi tudás nélkül is). Mivel az általunk létrehozott szótár (sok modern szöveggel ellentétben) jelöli a manysi magánhangzók hosszúságát (ami fontos a helyes kiejtéshez) a használók a helyes kiejtést is elsajátíthatják segítségével. A szó- és paradigmagenerátor segítségével nyelvtani tudásukat gyarapíthatják (anyanyelvi beszélő segítsége nélkül is). Ez azért is fontos, mivel manysi nyelvű nyomtatott szótárak a mindennapi életben nehezen elérhetőek, az anyanyelvi beszélők manapság pedig már kevesen vannak (és annál is kevesebben vesznek részt a manysi nyelv tanításában).

2.2.2.5. Északi manysi wordnet

A projekt során északi manysi Wordnet kiépítését is megkezdtük. A Wordnet egy lexikális adatbázis, amely a szavakat szinonimahalmazokban tárolja és ezen halmazok között logikai relációkat (pl. hiperonima) is tárol. Mivel minden jelentéses egységnek van egy nemzetközi azonosító kódja, így szótárként is lehet használni olyan nyelvpárokra, amelyekhez már készült WordNet. A projektben a BCS1 néven futó alapszókincs létrehozása történt meg, amely 1148 szinonimahalmazt tartalmaz összesen 2161 szóalakkal. A BCS1 manysi nyelvre ültetése a már korábban is említett szótárból történik automatikusan, kézi ellenőrzéssel és pótlással.

A wordnet építése során tapasztalt legnagyobb nehézséget az anyanyelvi beszélők alacsony száma, és a rituális medveműnyelv jelentette. A medve a manysi kultúrában kiemelkedően szent állat, tiszteletéhez a részletes tabunyelv is hozzátartozott. Mivel a hiedelem szerint a medve megérti az emberi beszédet, így a medve, testrészei, a medvével kapcsolatos minden tevékenység (különösen a medve vadászat) megnevezésére a beszélők tabukifejezéseket használnak, hogy a medve ne értse. Ennek eredményeként a medveműnyelvhez tartozó szókészleteket külön-külön fel kell tüntetni a manysi wordnetben. A manysi wordnet jelenleg mintegy 300 synsetből (szinonímahalmaz) áll, amelyben szófaji megoszlás szerint a főnevek dominálnak (210 névszó), ezt követi a mintegy 90 igei synset.

A wordnetet szinoníma szótárként, szókincs bővítésére, angol nyelvű megfelelő kifejezések keresésére tudják használni a nyelv beszélői.

A wordnetből minta megtekinthető a projekt honlapján itt: http://www.ieas-szeged.hu/finugrevita/results_hu.html

2.2.2.6. Déli manysi adatfeldolgozás

Kihalt dialektus lévén a déli manysi korpuszt nem kortárs szövegekből vagy interjúkból építettük, hanem két nyelvész, Munkácsi Bernát 1888-ban és 1889-ben, valamint Artturi Kannisto 1903-ban és 1904-ben, vagyis több mint száz évvel ezelőtt gyűjtött szövegmutatványait használtuk fel. Mindketten hosszú, átfogó, a nyelvi adatok mellett a mindennapi életre, folklórra, anyagi kultúrára, hiedelemvilágra is kiterjedő kutatómunkát folytattak, gyűjtésük megjelent anyagának a Vogul Népköltési Gyűjtemény negyedik, illetve a Wogulische Volksdichtung különböző köteteiben publikált déli manysi szövegei, főleg népmesék, énekek, találós kérdések szolgálnak a déli manysi korpusz alapjául. A latin betűs, diakritikus jelekkel gazdagon ellátott lejegyzésű szövegeket digitalizáltuk és elérhetővé tettük a korpusz honlapján (http://norbertszilagyi91.wixsite.com/tawdamansi). Az adatbázis 2400 mondatból és 11 500 szóalakból áll, ebből 5000 külön lexikális egység. A korpusznak SIL FieldWorks Language Explorerrel (FLEx: http://fieldworks.sil.org/flex/) kézzel annotált, morfológiailag elemzett verziója is elkészült, az annotáció a magyar fordítással együtt elérhető a honlapon, vagyis az adatok párhuzamos korpuszként is használhatók.

Mindegyik korpuszunk felhasználhatóságát bővítendő, az adatokat morfológiai és szintaktikai annotációval láttuk el. E célra az Univerzális Dependencia projekt által kifejlesztett morfológiai és szintaktikai címkekészletet alkalmaztuk, különös figyelmet fordítva a rokon nyelvekre (pl. finnre és magyarra) kidolgozott nyelvspecifikus jegyekre.

A 19. és 20. század fordulóján tevékenykedő kutatók manysi szóanyagából összeállított szótárak csak jelentős késéssel jelentek meg, a szótárak minden, így az azóta kihalt dialektusok szóanyagát is tartalmazzák. Legjobb tudomásunk szerint kizárólag a déli manysi szóanyagot tartalmazó szótár eddig nem készült, ezért a déli manysi korpusz adatai alapján déli manysi szószedetet hoztunk létre. A déli manysi szótár a korábban bemutatott korpuszban szereplő minden szót és toldalékot tartalmaz, annak magyar fordításával és morfológiai információkkal együtt. A lista az egyes elemek alapváltozatait, allomorfjait is tartalmazza.

Ennek a kihalt változatnak a digitális anyagú hozzáférése azért jelentős, mert ugyan a mai beszélők nyelvhasználatát kevésbé tudja segíteni, de szélesítheti a manysik ismereteit a nyelvükről és kultúrájukról, s ezért, áttételesen, növelheti presztízsüket. Nyelvészeti kutatás céljaira való felhasználhatósága az anyagnak azonban igen nagy.

3. Szociolingvisztikai vizsgálatok

3.1. Általános szociolingvisztikai keret

A modern technológia fejlődése, az internet és okostelefonok elterjedése lehetővé teszi azt, hogy az emberek a világ minden táján valós időben kommunikáljanak egymással.

Az emberek közti kommunikáció, illetve a gép–ember kommunikáció elősegítését szolgálják a nyelvtechnológiai eszközök és alkalmazások, mint például helyesírás-ellenőrzők, gépi fordítóoldalak vagy keresőprogramok, a digitális világban történő kommunikációt pedig különféle online erőforrások és alkalmazások segítik. Problémát jelent azonban az, hogy míg a világ nagy nyelveire jelenleg is számos nyelvtechnológiai eszköz létezik, addig a kisebbségi nyelvekre sokszor még a legalapvetőbb digitális nyelvi eszközök sem léteznek. Projektünk elsődleges célja volt, hogy olyan nyelvtechnológiai eszközöket készítsünk finnugor kisebbségi nyelvek beszélőinek számára, amelyek megkönnyítik számukra a digitális világban való anyanyelvi (kisebbségi nyelvi) kommunikációt.

A kisebbségi nyelvek nemcsak beszélőik számában különböznek más nyelvektől, hanem legfőképpen abban, hogy esetükben leginkább olyan nyelvekről van szó, amelyek nem hivatalosak országukban (hanem egy nagy, hivatalos státusszal rendelkező nyelv mellett, annak árnyékában léteznek), és beszélőik is ezért legtöbbször olyan kétnyelvűek, akik a hivatalos/többségi nyelven (végzik vagy) végezték iskolai tanulmányaikat, hivatalos és írott funkciókban, a munkahelyen leginkább azt használják. Ily módon a kisebbségi nyelv a privát szférára (családon belüli, barátok közötti stb.) és azon belül is a szóbeli kommunikációra korlátozódik, írásban kevéssé használatos lesz.

Napjainkban a digitális (azaz számítógépes közegű) nyelvhasználat (pl. e-mail írás és olvasás, chatelés, fórumozás, kommentelés, blogírás és -olvasás) megnövelte a nyelvhasználók írott nyelvhasználatát. Kétnyelvű beszélők esetében ezért elsőrendűen fontos kérdés, hogy tudják-e kisebbségi nyelvüket digitálisan használni. A felhasználói oldalról is hasznos nyelvtechnológiai alkalmazások létrehozásához, mint például a fentebb is említett helyesírás-ellenőrző vagy gépi fordító, elengedhetetlen, hogy rendelkezésre álljanak az alapszintű nyelvfeldolgozó technológiák az adott nyelvre. Projektünk számára nem kevésbé volt lényeges a készített nyelvtechnológiai eszközök sikerességének vizsgálata, ezeket a szociolingvisztikai méréseket az eszközök kifejlesztése előtt és után is terveztük elvégezni az eredmények összehasonlíthatósága érdekében.

3.2. Kísérleti szociolingvisztikai vizsgálat: a számik

A projekt egyik első lépése volt a már létező számi nyelvtechnológiai eszközök használóinak körében végzett online felmérés.

A számi (ami valójában nem egyetlen nyelv; itt elsősorban az északi számit jelöli a megnevezés) mint négy országban beszélt kisebbségi nyelv mára a skandináv országokban és Finnországban megkapta a regionálisan hivatalos nyelv státuszát, Oroszországban azonban a helyzet jóval nehezebb. Norvégia, Svédország és Finnország kormányai anyagilag is segítik a számi nyelvek fennmaradását, revitalizációját. A számi beszélők hatalmas területen, több mint 400 000 négyzetkilométeren elszórva élnek, összesen kb. 70-80 ezren, ma pedig már igen sokan elköltöztek délebbi területekre, vagyis nagy számú ún. city-számiról is beszélhetünk. Pontos számot sem az etnikai számik, sem a számi beszélők esetében nem lehet megállapítani, aminek elsődleges oka az, hogy a népszámlálás során a nemzetiségi hovatartozás önbevalláson alapul, és sok számi felmenővel rendelkező ember inkább más etnikumúnak mondja magát. Másrészt pedig a számi nyelvismeret, nyelvtudás és az anyanyelv (első nyelv) nem feltétlenül esik egybe, mivel nagyon sokan csak felnőttként tanulnak meg számiul, s ilyenkor gyakori, hogy nem a családban korábban használt

változatot, hanem egy magasabb presztízsű, kevésbé veszélyeztetett számi nyelvet választanak, legtöbbször az északi számit (amelyre a legtöbb digitális eszközt is kifejlesztették). A számik anyanyelvhasználatának fontos eleme, hogy a nyelvük digitálisan is hozzáférhető, a Giellatekno és a Divvun oldalai számos nyelvtechnológiai eszközzel segítik a beszélőket (http://giellatekno.uit.no, http://divvun.no).

Online kérdőívünkben a kérdések részletesen felmérték egyrészt a felhasználókat, másrészt a használt eszközökkel kapcsolatos véleményüket, illetve a fejlesztésekre vonatkozó elvárásaikat.

A kérdőív a http://www.ieas-szeged.hu/finugrevita/survey.html linken keresztül, négy nyelven (angolul, finnül, norvégul és svédül) volt kitölthető: https://docs.google.com/forms/d/e/1FAIpQLScCkIkIzpFQ5cUwRjx_Tk1JOk5-zzU2Xud79gs_-Vhqh_j0CQ/viewform, https://docs.google.com/forms/d/e/1FAIpQLSeAnmRQXUtVNSK5hB6ob8zyVnWH0CYz4FHLs4a0fl0GtOIuGQ/viewform, https://docs.google.com/forms/d/e/1FAIpQLSfLxPb0l_cfwec5jme5yt8XyFOShUFmY5ZC2imniToj8_MV5g/viewform, https://docs.google.com/forms/d/e/1FAIpQLSddhv85q2uOfw5U3nGYURN7mgixQddJdeavSIffK9PFfpi07A/viewform. A számi azért nem szerepelt ezen nyelvek között, mert a kutatócsoportunk nyelvi kompetenciáját meghaladta volna a számi fordítás, ráadásul több számi nyelven (az északin kívül legalább déli, lulei, inari és kolta számin) kellett volna elkészíteni a kérdőívet, miközben a válaszadók mindegyike jól beszéli a fenti nyelvek valamelyikét.

Az első kérdéscsoport a válaszadók személyére vonatkozott: rákérdeztünk az anyanyelvükre, számi nyelvismeretükre, illetve az egyéb beszélt nyelveikre, megkérdeztük, mely országban élnek, mi a legmagasabb iskolai végzettségük, milyen neműek, mi a foglalkozásuk, stb.

A második kérdéskör a használt alkalmazások és a meglátogatott oldalak iránt érdeklődött. Megkérdeztük, hol hallottak először a felmérésben szereplő nyelvtechnológiai eszközökről, valamint a Giellatekno és a Divvun 2014-ben létező minden programjára (szótárak, elemzők, nyelvoktató programok, billentyűzet-kiosztás, nyelvi játékok stb.) külön-külön rákérdeztünk, hogy használja-e a válaszadó, és ha igen, mennyire elégedett vele.

Végül a harmadik kérdéstípus a további igényeket mérte fel: mit szeretnének még látni a felhasználók, például korpuszbővítés esetén milyen típusú szövegeket tartanának fontosnak, milyen további szótárak és fordítóprogramok lennének szerintük hasznosak és fontosak (itt a különböző nyelvpárokat is felmértük), min szeretnének változtatni a már meglévő alkalmazásokban és programokban.

A felmérés 2014 végén indult, majd először 2015 márciusában, később még egyszer 2015 nyarán összesítettük a kapott adatokat. Az eredményeket az oului Finnugor Kongresszuson előadás keretében mutattuk be 2015-ben. Az érdeklődés igen nagy volt, a Giellatekno és a Divvun (a két számi nyelvtechnológiai oldal) fejlesztői már akkor felvetették, hogy ennél részletesebben is szeretnének megismerkedni az adatokkal. A felmérésünk által felszínre hozott felhasználói kérések közül a fejlesztők azóta nem egyet már teljesítettek is, így pl. a korábbinál több digitális szótár található meg az oldalakon, a

nyelvoktató és fordítóprogramokat fejlesztették, a már előzőleg is meglévő, de csak nehezen kereshető elemző- és szótári programokban is több változást is bevezettek. Mindennek eredményeképpen a legtöbb nyelvtechnológiai eszköz könnyebben hozzáférhető és felhasználóbarát formában van jelen az interneten.

2018 januárjában került sor Tromsøben az egyetem, ill. a számi nyelvtechnológusok meghívására a felmérés eredményeinek és az azóta eltelt időszakban megvalósított fejlesztéseknek a részletes, személyes megbeszélésére. Mindkét portál üzemeltetői alaposan belemerültek a felhasználói vélemények és igények elemzésébe, meghallgatták a mi észrevételeinket és tanácsainkat is. Mind a Giellatekno, mind a Divvun kilátásba helyezett olyan megoldásokat, amelyek tovább egyszerűsítik a szolgáltatások használatát, és bővítik a lehetséges programok körét.

Külön problémát okozott, hogy a válaszadók száma nem volt túl magas, mindössze 67 adatközlő véleményére támaszkodhattunk, akik közül a legtöbben maguk is kutatók, nyelvészek. Ugyanakkor egészen biztosak vagyunk abban is, hogy ennél jóval többen használják ezeket az oldalakat és szolgáltatásaikat a nem szakemberek közül is, azonban a hétköznapi, számi nyelvet a mindennapos életben használó beszélők viszonylag kevesen töltötték ki. Ennek oka leginkább az lehet, hogy a számikban még mindig működik egyfajta félelemmel vegyes távolságtartás, ha nem számi nemzetiségű személy akar foglalkozni bármiféle számi üggyel. A nem is nagyon távoli múltban megvetéssel néztek a számikra még a skandináv országokban is, az iskolákban büntetés járt a számi nyelv használatáért. Mára a helyzet sokban megváltozott, de az évszázados beidegződés nem múlt el. A felmérés indulásakor minden számi nyelvvel foglalkozó vagy számi intézménynek elküldtük a linket egy kísérőlevéllel, ennek ellenére a nyelvhasználók közül igen kevesen töltötték ki a kérdőívet. A másik ok pedig talán az volt, hogy a kérdőív a két említett oldal minden eszközére rákérdezett, a felhasználók pedig nem mindegyiket ismerik, csupán azokat, amelyekre szükségük van, ezért a kitöltés során találkozhattak számukra irreleváns kérdésekkel is, és ez megzavarhatta őket.

A felmérés mégis nagyon hasznosnak bizonyult a Giellatekno és a Divvun fejlesztői számára, hiszen ők nincsenek közvetlen kapcsolatban a felhasználókkal, nem érkeznek hozzájuk visszajelzések az eszközeikkel kapcsolatban, csak azt tudják lekérdezni a szerver adataiból, hogy mennyien és milyen rendszerességgel látogatják az oldalaikat. Az általunk végzett felmérés azonban megmutatta nekik is, mire van leginkább szükségük a számi beszélőknek és felhasználóknak (a kutatók szempontjai és igényei ebben kevésbé fontosak, ők akár önállóan is közlik a véleményüket), viszont a kívülállók, a hétköznapi nyelvhasználók véleménye csak ilyen formában tud utat találni a nyelvtechnológusokhoz. Éppen ezért a tromsøiek felvetették a további együttműködés lehetőségét is, hogy készítsünk egy újabb, a mai helyzetet körüljáró felmérést is, amelynek eredményeit szintén beépíthetik saját munkájukba. Mivel a FinugRevita projekt a végére érkezett, résztvevői pedig különböző intézmények alkalmazottai, ez az együttműködés egyelőre nehezen megvalósítható, de reméljük, hogy sikerül valamilyen formában folytatni a közös munkát.

3.3. Digitális nyelvhasználat felmérése: kérdőív

A számi kérdőív kialakítása, terjesztése, és a beérkezett válaszok kiértékelése során szerzett tapasztalatokat is felhasználva a digitális nyelvhasználat felmérésére kérdőívet

dolgoztunk ki, amellyel udmurt és manysi beszélők körében szándékoztunk adatot gyűjteni. A kérdőív részben támaszkodik az ELDIA Projektben kidolgozott European Language Vitality Barometer (EuLaViBar) kérdőívének (Spiliopoulou Åkermark et al. 2013) egyes részeire. A EuLaViBar kérdőív veszélyeztetett kisebbségi nyelvek vitalitásának felmérésére szolgáló kutatói eszközrendszer egyik alkotóeleme, amelyet készítői azzal a szándékkal hoztak létre, hogy ilyen nyelvekkel foglalkozó nyelvészek részére egységes módszertani eszközcsomagot (kérdőívet, elemzési és statisztikai útmutatót) bocsássanak közre. Mivel a EuLaViBar kérdőív a mi céljainkra egyes részeiben túl részletes és hosszú (pl. az általános nyelvhasználati szokások felmérése terén), más téren pedig nem elég részletes, ezért azt saját céljainknak megfelelően átdolgoztuk: az általános részét lerövidítettük, a digitális nyelvhasználattal kapcsolatos részt pedig számos kérdéssel kiegészítettük. Az így kapott kérdőívnek kb. egyharmada saját fejlesztésünk eredménye. A kérdőívet a felméréshez lefordítottuk oroszra, illetve lefordíttattuk udmurt nyelvre is, és online kitölthető módon készítettük a Google rendszerén belül. A manysi nyelvhasználók számára készített változatot csak oroszul, az udmurt nyelvhasználók számára készített változatot oroszul és udmurtul készítettük el.

A kérdőív személyes adatok (nem, életkor, születési hely, iskolai végzettség, foglalkozás) után az adatközlő által beszélt nyelvekre kérdez (anyanyelv, egyéb nyelvek, ezek tudásának szintje, elsajátításának módja, családtagokkal használt nyelvek, különböző nyelvhasználati színtereken használt nyelvek), majd a nyelvi attitűdökre, valamint az udmurt verzióban az udmurt nyelvhasználattal kapcsolatos törvényi keretekkel kapcsolatos tudásra és véleményre (ez a rész a manysi nyelvre vonatkozóan nem volt releváns, mivel a Hanti-Manysi Autonóm Körzetben a manysi nyelv nem rendelkezik semmiféle törvényileg szabályozott pozícióval).

A kérdőív számunkra legfontosabb része a digitális nyelvhasználattal kapcsolatosan tesz fel kérdéseket az adatközlőknek: ezeket a kérdéseket a digitális térben való létezésben legnagyobb tapasztalattal rendelkező fiatal (PhD hallgató) kollegáink javaslataira alapozva dolgoztuk ki. Ez az internethasználattal és aktív online nyelvhasználattal kapcsolatos kérdéseket tartalmazó rész kitér a (hivatalos állami és egyéb) internetes tartalmakhoz, kisebbségi nyelven létező számítógépes szoftverekhez való hozzáférésre, egyéni emailezési, sms-ezési, chatelési szokásokra, blogok, fórumok olvasására és azokhoz való hozzászólások írására, egyéb online tevékenységekre (oldal vagy online csoport gondozása, mémek, rövid képregények gyártása, fotók és videók megosztása, online játékok játszása, Google keresésekre), valamint ugyanazokkal a kommunikációs partnerekkel való online és offline kommunkációban fellelhető nyelvválasztási és nyelvhasználati különbségekre.

Kérdőíveink megtalálhatók itt:
Manysi kérdőív (oroszul):
https://docs.google.com/forms/d/e/1FAIpQLSc5ywkP-V4xBKXtD5sqxqWkLelY2_dOXNLi67o2-1A2j2238Q/viewform
Udmurt kérdőív (oroszul):
https://docs.google.com/forms/d/e/1FAIpQLScZduGKLABgk69Ij1ZAB16PZEe8AIbFxdfa14SIHyim8sokAw/viewform
Udmurt kérdőív (udmurtul):

https://docs.google.com/forms/d/e/1FAIpQLSe9rmPXov-vYEF0LqMlG9t0QrgF3eBbp4IMHZn0sumBnqd64w/viewform

A kérdőívekből később rövid(ebb) orosz nyelvű verziót is készítettünk abban a reményben, hogy annak kitöltésére talán nagyobb hajlandóságot mutatnak adatközlőink. E rövidebb kérdőív manysi nyelvre vonatkozó verzióját itt találhatjuk: https://docs.google.com/forms/d/e/1FAIpQLScCpn9_Evdyv6I1AmRpqrRpTbcdnZiKQGcnlvHPk-2b0Huvlw/viewform

Bár a kérdőíves szociolingvisztikai felmérésünk az adatközlők válaszadó kedvének hiánya miatt végül nem történt meg, véleményünk szerint a kérdőív jól felhasználható más kisebbségi nyelvek beszélői között végzendő, digitális nyelvhasználatot feltérképező vizsgálatokhoz, így az általunk létrehozott kérdőívet mindenképpen a projekt által felmutatható eredménynek tartjuk.

3.4. Manysi szociolingvisztikai vizsgálatok

A fent részletesen tárgyalt kérdőívet manysi beszélők körében online és a 2015-ös terepmunka során személyesen is megpróbáltuk kitöltetni, illetve interjú formájában lekérdezni.

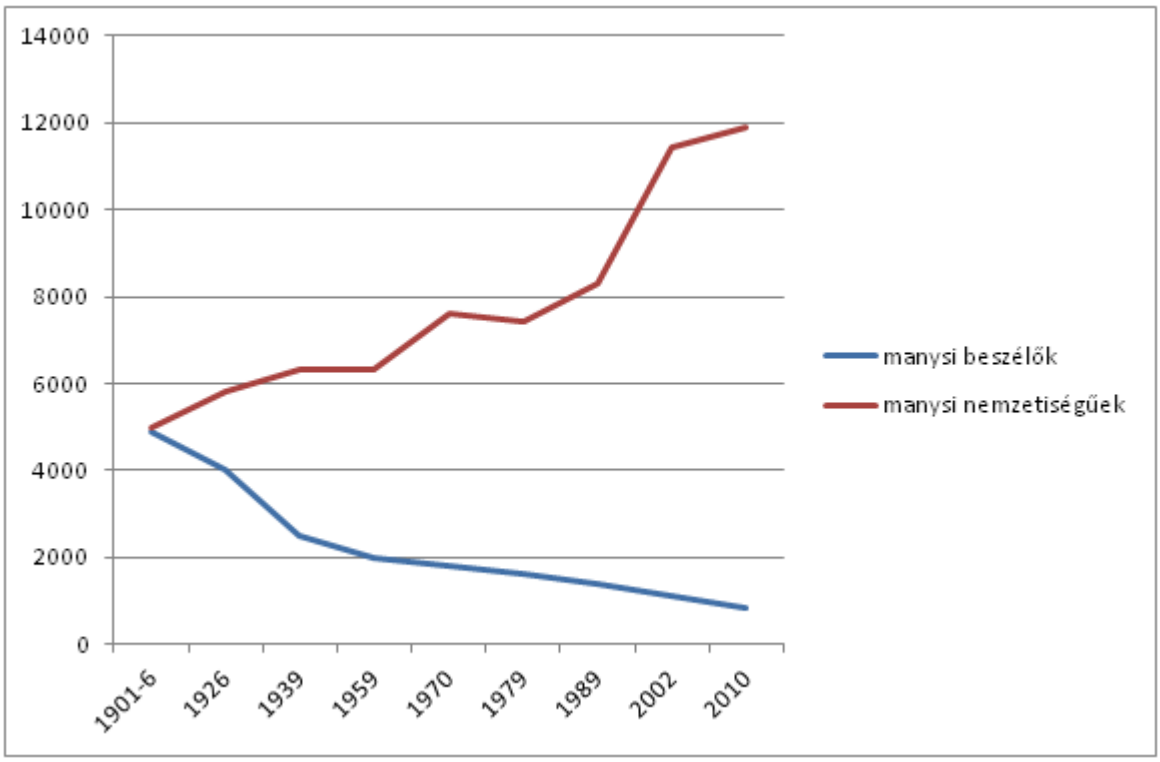
A nyelvhasználók válaszadási hajlandósága jelentősen alulmúlta az előzetes várakozásainkat: míg az online kérdőívet alig néhány ember (n=19) töltötte ki, a terepmunka során erre egyetlen adatközlő sem vállalkozott. A távolságtartásra magyarázatot adhat egyrészt a részletes kérdőív hossza (41 főkérdés további alkérdésekkel), másrészt pedig az érzékeny témát érintő kérdéseket feltevő idegen személyek iránti gyanakvás, félelem. (A mai Oroszországban alacsony presztízsű kisebbségi nyelv, főként pedig őshonos nyelv beszélőjének, ilyen kisebbségi közösség tagjának lenni a hivatalos állami politika fényében nem sok pozitívummal kecsegtető helyzet, amelyet a beszélők leginkább privát helyzetként élnek meg, és ezért kevéssé akarnak megnyilatkozni róla, akár anonim formában is.)

Nyelvtudást, nyelvhasználatot - jóval átfogóbb kérdések alapján - vizsgáló adatsort csak az összoroszországi népszámlálások során gyűjtöttek eddig a manysik körében. Ezek az adatgyűjtések nem voltak önkéntesek, ugyanakkor épp részben kötelező voltuk miatt, részben nem egyértelmű terminológiájuknak köszönhetően eredményeik sem teljes mértékben megbízhatóak. Az általunk kidolgozott kérdőív nyelvhasználók általi kitöltését természetesen semmilyen körülmény nem kényszerítette, ugyanakkor - mivel a nyelvtudás és a nyelvhasználat az egyik legfontosabb identitásképző elem - a nyelvhasználatra vonatkozó kérdések a veszélyeztetett nyelvi környezetben különösen érzékeny témának bizonyultak.

A második szociolingvisztikai felmérést a korábbi tapasztalatok tükrében némiképp másképp terveztük meg: a kérdőív egy rövidebb és lényegretörőbb változatát kidolgoztuk, de készülve arra a lehetőségre, hogy ugyanúgy alacsony lesz a kitöltési kedv, tervbe vettük a jelenlegi nyelvhasználati területek szisztematikus végiglátogatását és a nyelvtudási, nyelvhasználati szokások empirikus módon történő felmérését. Ennek a résztvevő megfigyeléssel és interjúhelyzetben végzett adatgyűjtésnek az eredményei a következőképpen foglalhatók össze.

Oroszország multikulturális állam, melynek területén több mint 200 etnikum él, a szabad népességmozgás hatására a homogén etnikumú területek eltűnőben vannak, már csak

mikroszinten kimutathatóak, mint például a manysi (időszakosan lakott) falvak esetében. A különböző nemzetiségek az ország népességének ötödét teszik ki, ezért az orosz népesség túlsúlya folytán a kulturális és nyelvi hatások a ruszifikációs folyamatokat erősítik. Az állam- és közvetítőnyelv az orosz, és a kulturális normák is az oroszhoz igazodnak. Az etnikumokra, amennyiben azok saját nyelvvel bírnak, jellemző a diglosszia, ahol az *emelkedett* nyelv az orosz, míg a *közönséges* nyelv az adott nemzetiség nyelve. Az eloroszosodás és nyelvvesztés miatt viszont már nem állítható kivétel nélkül mindegyikre, hogy a teljes népcsoport jellemzően abszolút két- vagy többnyelvű. A 2010-es népszámlálási adatok alapján a magukat manysi nemzetiségűnek vallók a körzet lakosságának 9%-át teszik ki, míg a közülük manysi nyelvi tudásról nyilatkozók az össznépesség 0,054 %-a (kevesebb mint egy ezreléke).



1. táblázat. A manysi beszélők és magukat manysi nemzetiségűnek vallók száma a 20. század kezdetétől napjainkig [forrás: Kannisto-Nevalainen 1969; Oroszországi népszámlálás 2010; Rombangyejeva; Sipőcz–Dolovai 2001]

A 2010-es népszámlálási adatok alapján Oroszország területén 11 873-an nyilatkoztak manysi nemzetiségükről, közülük 1773 „anyanyelvként" jelölte meg a manysit, jóllehet itt az orosz *родной язык* kifejezés a nemzetiség nyelvét jelenti nem kizárva, hogy a beszélő tényleges anyanyelve más. A nyelvtudást kérdező népszámlálási ívek adatai szerint majd mindannyian tudnak oroszul, a nagy világnyelveken kívül 834-en manysiul, 37-en komiul, 24-en hantiul, 6 mari, 5 magyar, 4 nyenyec, 3 udmurt és 1 finn nyelven is tudó manysi nemzetiségű van, amennyiben csak a finnugor nyelveket vesszük szemügyre.

	Tyumenyi Terület (összefoglaló nagyobb közig. ter.)						Szverdlovszki Terület		más körzetek	
	Hanti-Manysi Autonóm Körzet - Jugra		Tyumenyi Terület		Jamal-Nyenyec Autonóm Körzet					
	városi	falusi	városi	falusi	városi	falusi	városi	falusi	városi	falusi
0-19	2412	1846	104	14	32	10	39	49	-	-
20-39	2121	1533	162	16	37	17	43	37		
40-59	1244	1232	93	38	40	20	33	31		
60+	303	286	34	10	6	4	10	9		
∑	6080	4897	393	78	115	51	125	126	5	3
	10977		471		166		251		8	
	11873									

2. táblázat. A manysi nemzetiségű lakosság száma a közigazgatási terület és település típusa szerinti bontásban. [forrás: http://www.gks.ru/ Oroszországi népszámlálás 2010 év = *Всероссийская перепись населения 2010 года.* Федеральная служба государтсвенной статистики РФ].

A népszámlálási adatok alapján közigazgatási egység, lakhely típusa és életkor szerinti bontásban is elemezhetjük a manysi nemzetiségűek jellemző adatait. A statisztika 5 éves bontásban közli a számokat, viszont a generációk könnyebb összehasonlíthatósága miatt azokat 20 éves generációkba csoportosítottuk, így gyermek (0-19 év), fiatal felnőtt (20-39 év), középkorú (40-59 év) és idős (60+) korosztályokat különítettünk el. Lévén közel egy évtizede készült a népszámlálás, és mivel a tapasztalatok alapján feltételezzük, hogy a manysiul beszélők legfőképp a középkorú és nagyobb számban az idős korosztályban találhatók, így a 2010 óta bekövetkezett természetes fogyás következtében arányosan kevesebb anyanyelvi beszélővel számolhatunk. A vidéki falvakból az egyre gyorsabban fejlődő városokba költöző népességgel a fiatalok száma kétszerese a városokban mint a falvakban, míg a középkorú és idős korosztályokban a településtípus szerint fele-fele a népesség eloszlása. A központosítottabbá váló oktatásnak köszönhetően az eddig iskolával rendelkező falvakban és községekben elvégzett általános iskola után a középiskolai vagy felsőoktatásért az ideiglenesen a nagyobb városokba vándorló fiatalok nagy hányada ott is

telepszik le. Ennek hatása a kis falvak elnéptelenedése és a fiatalabb generációk nagyobbrészt manysi közegből multikulturális orosz nagyvárosi közegbe való kerülése és asszimilációjuk a modern életkörülmények alkotta kihívásokhoz. A Kannisto által összeállított *Statistik über die Wogulen* adatai alapján homogén manysi falvak legtöbb esetben az elvándorlások miatt elnéptelenedtek, vagy a betelepülő más nemzetiségek miatt heterogén kulturális és nyelvi közegek jöttek létre, mely utóbbi a gazdaságilag kiemelt helyen fekvő falvakra jellemző, ahol a 20. század vége felé felgyorsuló nyersanyag-kitermelés határozza meg kizárólagosan a lehetőségeket.

A 2017-es terepmunka tapasztalatai alapján a helyi manysi nemzetiségűek kapcsolati hálójának és közösségeinek megismerése révén szubjektív szempontok (beszélők önbevallása és a terepmunkás tapasztalatai) alapján a társalgási és anyanyelvihez közelítő (egységes referenciakeret szerinti B2-C2 nyelvi szintekbe sorolható) kompetenciával rendelkező manysi beszélők számát a következő táblázatban szemléltetjük korosztályok szerinti bontásban.

	65 -	**40 - 65**	**20 - 40**	**0 - 20**
obi	<5	10-20	0 ?	0
alsó-szoszvai	<10	25-75	0 ?	0
közép-szoszvai	<25	200-400	10-20	0
felső-szoszvai	<5	10-20	20-50	5-10
felső-lozvai	<5	5-10	10-20	5-10
jukondai	<3	0	0	0

3. táblázat. A manysi nyelvet anyanyelvihez közelítő szinten beszélők számának hozzávetőleges becslése önbevallás, terepmunkás saját tapasztalatai és napjaink nyomtatott tudományos és sajtószövegeinek adatai alapján [saját adatok]

A beszélői korfa és a beszélők lakhelye alapján kikövetkezhethető tendenciáknak megfelelően a 2015-ös és a 2017-es terepmunkák során kapott visszajelzések szerint egyaránt a nyelvelsajátítást, a nem anyanyelvi kompetenciájú beszélők nyelvhasználatát segítő nyelvtechnológiai eszközök, ezen belül is mindenekelőtt az online szótár, illetve a helyesírásellenőrző iránt a legnagyobb az érdeklődés. Mivel a legfrissebb manysi szótárat is több mint tíz évvel ezelőtt adták ki, alacsony példányszámban, üzletekben nem volt megvásárolható, így sem az iskolások, sem az érdeklődök nem juthattak (vagy csak nagy nehézségek árán) saját példányhoz, ezért az online szótár rég fájó hiányt pótol.

2017. június 21 és augusztus 14 között végzett terepmunkaút során az oktyabrszki körzetbeli az Ob folyó menti Vezsakori és Nyizsnyije Narikari falvakba, valamint a berjozovói körzetbeli Szoszva folyó menti Berjozovo, Igrim, Anyajeva és a Ljapin-folyó (Szigva) menti Sekurja, Szaranpaul, Jaszunt, Hurumpaul és Hoslog falvakban gyűjtött munkatársunk nyelvi hangzó anyagot (digitális hang- és videofelvételek). A kutatás ideje

alatt 20 adatközlővel rögzített hanganyag körülbelül 13 órányi nyers hangfelvétel, melynek nagy része manysi nyelvű beszélővel készített manysi nyelvű interjú. A hanganyaggal átfedésben lévő 9 órányi videofelvétel is nagyobb mennyiségű nyelvi adattal szolgál.

Ezen terepmunkaút során munkatársunk a manysi nyelv valós színtereire is eljutott. Összesen 15 fő 65 éven felüli, a manysi nyelvet folyékonyan beszélő adatközlővel készített szociolingvisztikai interjút (amelyekről digitális hangfelvételt is készített, összesen több mint 20 óra terjedelemben, valamint némelyikről videofelvételt is tudott készíteni, az adatközlő engedélye függvényében). A hitélethez köthető helyeken (szentház, temető, kegyhely stb.) tartott hitéleti eseményeken (elhunytról való megemlékezés, különböző áldozat bemutatások stb.), nagycsaládi összejöveteleken, manysi ünnepségeken vett részt, valamint a beszélők otthoni nyelvhasználatát is dokumentálta. Így olyan élethelyzetben találkozott manysi beszélőkkel, amely helyzetek egyrészt nagyon ritkán fordulnak elő (mivel a beszélők elszármaztak máshová), és kulturálisan szabályozott módon történhetnek (pl. temetőlátogatás a halott elhunyta után csak ötéven belül megengedett és lehetséges).

A terepmunka nyelve - a mai finnugrisztikai kutatásban egyedülálló módon - a manysi volt, ami lehetőséget adott arra, hogy az interjúk készítése és a nyelvi dokumentáció során a beszélők tényleges nyelvi kompetenciáját is felmérje szakmai megalapozottságú szubjektív tapasztalatai alapján, és valós képet kapjon a manysi nyelv mindennapi használatáról (így a beszélők viszonylagos számáról, egymással ápolt kapcsolatairól, a nyelvhez köthető viszonyáról, és a nyelvi revitalizációhoz fűződő gondolatairól, szükségleteiről).

Ez az anyag jelenleg transkripcióra és annotációra vár, ami után spontán hangzó nyelvi korpusz építhető belőle (párbeszédekkel, monológokkal) és alapjául szolgálhat fonetikai, morfológiai, szintaktikai, pragmatikai és diskurzus elemzés jellegű vizsgálatoknak. Ennek legfőbb jelentősége, hogy hangzó nyelvi korpusz eddig még nem készült (főleg nem spontán nyelvi korpusz, és nem több beszélőtől), így ez az anyag kiváló kiindulópontja lehet az írott nyelvi korpusszal való összehasonlításnak. Mai manysi nyelvet szabadon beszélő beszélőktől származó anyag ilyen nagy mennyiségben eddig nem volt rögzítve és elemezve.

3.5. Az udmurtiai szociolingvisztikai vizsgálatok

Az udmurt beszélők számára is elkészült a magyar-udmurt nyelvtechnológiai eszköz elkészítése előtt lekérdezendő kérdéssor, melyet online töltettünk ki. A válaszadási hajlandóság ebben az esetben is jelentősen alulmúlta az előzetes várakozásainkat: az online kérdőívet alig néhány ember töltötte ki.

Az udmurt esetében a fejlesztett számítógépes eszköz a magyar-udmurt-magyar digitális szótárunk lett. A vele összekapcsolt korpusz fejlesztése elmaradt, mivel a pályázat beadása és a kutatás megkezdése után lett nyilvános és ismert a Timofey Arkhangelskiy vezette moszkvai kutatócsoport által fejlesztett Udmurt Corpus (http://web-corpora.net/UdmurtCorpus/search/). Mivel az általunk fejlesztett szótár egy szűk, azonban az udmurt nyelv alakulása szempontjából rendkívül fontos csoportot ér el, az Udmurt Állami Egyetemen udmurt nyelv és irodalom szakon tanuló, leendő udmurt tanárokat és értelmiségieket. Körükben a második szociolingvisztikai felmérés elvégzése értelmetlen lett volna, mivel egyrészt a használók jelentős részével gyakorlatilag napi munkakapcsolatban

vannak pályázatunk munkatársai, másrészt a kapott eredmények az udmurt nyelvhasználók egészére nézve nem szolgáltak volna használható adatokkal.

4. Összefoglalás

Projektünk elsődleges céljait, a számítógépes nyelvi eszközök létrehozását veszélyeztetett oroszországi finnugor nyelvekre – a manysira és az udmurtra – teljesítettük, hiszen létrehoztunk elektronikus udmurt–magyar–udmurt szótárat, elektronikus manysi–magyar és manysi–orosz szótárat, manysi korpuszt, manysi helyesírás elemzőt, morfológiai elemzőt, és wordnetet. A másodlagos cél, a szociolingvisztikai felmérések elkészítése, leküzdhetetlen nehézségekbe ütközött, ezért nem valósult meg. Helyette azonban a projekt egyik résztvevője páratlan manysi nyelvi anyagot tudott gyűjteni, amely alapján a jövőben lehetőség nyílik majd spontán beszélt manysira alapuló hangzó nyelvi korpusz létrehozására (párbeszédekkel, monológokkal) és ezen alapuló fonetikai, morfológiai, szintaktikai, pragmatikai és diskurzus elemzés jellegű vizsgálatokra.

Hivatkozások:

Kannisto, Artturi. 2013 *Wogulisches Wörterbuch.* Helsinki: Kotimaisten Kielten Keskuksen Julkaisuja

Kannisto, Artturi - Nevalainen, Jorma. 1969. Statistik über die Wogulen. Gesammelt von Artturi Kannisto. Bearbeitet und herausgegeben von Jorma Nevalainen. Suomalais-Ugrilaisen Seuran Aikakauskirja - *Journal de la Société Finno-Ougrienne* **70.** 1970, Helsinki, Suomalais-ugrilainen Seura. 4.: 1-95.

Kozmács István. 2002. *Udmurt-magyar szótár*. Szombathely: Savaria University Press Alapítvány.

Munkácsi, Bernát - Kálmán, Béla. 1986. *Wogulisches Wörterbuch.* Budapest: Akadémiai Kiadó.

Nagy Zoltán. 2015. Szibéria néprajza és a város: Akik kimaradtak az összefoglalókból. In: Szeverényi Sándor és Szécsényi Tibor, szerk., *Érdekes nyelvészet.* Szeged: JATEPress, 57-72.

Oroszországi népszámlálás 2010 = *Всероссийская перепись населения 2010 года.* Федеральная служба государтсвенной статистики РФ

— 20. Владение языками населением коренных малочисленных народов Российской Федерации http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol4/pub-04-20.pdf

— 22. Население коренных малочисленных народов Российской Федерации по родному языку http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol4/pub-04-22.pdf

— 25. Население коренных малочисленных народов Российской Федерации по территориям преимущественного проживания, возрастным группам и полу http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol4/pub-04-25.pdf

Riese, Timothy. 2001. *Vogul.* Number 158 in Languages of the World/Materials. Munchen - New Castle: Lincom Europa.

Rombangyejeva-Kuzakova 1982 = Ромбандеева, Е. И. - Кузакова, Е. А. 1982 *Словарь мансийско-русский и русско-мансийский.* Ленинград: Просвещение.

Rombangyejeva 1973 = Ромбандеева, Е. И. 1973. *Мансийский (вогулский) язык*. Москва**:** Наука.

Rombangyejeva 1976 = Ромбандеева, Е. И. (1976): Мансийский язык. In: *Основы финно-угорского языкознания: марийский, пермские и угорские языки.* Наука, Москва: 229-239.

Rombangyejeva 2005 = Ромбандеева, Е. И. 2005. *Русско-мансийский словарь* Санкт-Петербург: Миралл .

Sipőcz Katalin - Dolovai Dorottya (2001): A vogulok (manysik). In: Csepregi Márta (szerk.): *Finnugor kalauz.* Budapest: Panoráma, 48–59.

Spiliopoulou Åkermark, Sia, Johanna Laakso, Anneli Sarhimaa, Reetta Toivanen, Eva Kühhirt, and Kari Djerf. 2013. ELDIA EuLaViBar. <http://www.eldia-project.org/index.php/eulavibar>.

Thieberger, N. and A. L. Berez. 2012. Linguistic data management. In: N. Thieberger, editor, *The Oxford handbook of linguistic fieldwork*. Oxford: Oxford University Press, 90-118.