FinUgRevita: nyelvtechnológiai eszközök fejlesztése kisebbségi finnugor nyelvekre

Horváth Csilla[1], Kozmács István[2], Szilágyi Norbert[2], Vincze Veronika[3], Nagy Ágoston[1], Bogár Edit[1], Fenyvesi Anna[1]

[1]Szegedi Tudományegyetem, Angol-Amerikai Intézet
[2]Szegedi Tudományegyetem, Finnugor Tanszék
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
e-mail:finugrevita@gmail.com

Kivonat A jelenleg is futó nemzetközi FinUgRevita projekt (2013-2017) keretében olyan nyelvtechnológiai eszközöket fejlesztünk, melyek a kis finnugor népek közülük a manysik (vogulok) és udmurtok (votjákok) nyelvének digitális és online jelenlétét teszi lehetővé, és segíti az anyanyelvi beszélőket és a tanulni vágyókat a nyelvi kommunikáció különféle színterein. A kezdeti fázisban a két nyelv jelenkori leíró nyelvtanai alapján nyelvtani kivonatok készülnek, melyek a készülő online morfológiai elemző szabályrendszerét adják, míg az eddig megjelent nyomtatott szótárak szkennelésével, OCR-es elemzésével és manuális javítással az udmurt esetében 13000, míg a manysi esetében előreláthatólag 10-15000 szavas elektronikusan felhasználható szótár készül. A morfológiai elemző futtatásához és további nyelvtechnológiai eszközök fejlesztéséhez az interneten szabadon elérhető udmurt és manysi nyelvű tartalmakból nagy tokenszámú korpuszt építünk. A projekt célja, hogy a készülő eszközök online szabadon hozzáférhetőek legyenek az udmurt és manysi nyelvek beszélőinek és tanulóinak számára, és nem utolsó sorban kutatási célokra is alkalmazhatóak legyenek.

Kulcsszavak: udmurt, manysi, nyelvtechnológiai eszközök, veszélyeztetett nyelvek

1. Bevezetés

A modern technológia fejlődése, az internet és okostelefonok elterjedése lehetővé teszi azt, hogy az emberek a világ minden tájján 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 elő. 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. A projekt elsődleges célja, hogy olyan nyelvtechnológiai eszközöket készítsünk finnugor kisebbségi

nyelvek beszélőinek számára, amelyek megkönnyítik számukra a digitális világban való anyanyelvi 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 nyelvhasználó írott nyelvhasználatát. Kétnyelvű beszélők esetében ezért elsőrendűen fontos kérdés, hogy tudják-e kisebbségi nyelvüket digitálisan használni [1].

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. A kisebbségi nyelvek esetében a szövegfeldolgozás akár a karakterkódolás szintjén is problematikus lehet, amennyiben nem létezik egy egységesített (sztenderdizált), széles körben elterjedt karakterkészlet. A nyelvtechnológiai alkalmazások létrehozásához szükséges továbbá egy szegmentáló (mondatra, illetve azokat szavakra bontó) eszköz, morfológiai elemző és szófaji egyértelműsítő, a szövegek jelentésének megértésében pedig a szintaktikai és szemantikai mélyelemzők játszanak fontos szerepet. Ezen alaptechnológiák kifejlesztése egymásra épül: például a szegmentáló kimenetéhez, azaz az egyes szavakhoz rendel elemzést a morfológiai elemző, majd a szintaktikai elemző a szófaji kódokat is figyelembe véve elemzi a mondatokat stb.

A projekt elsődleges céljának eléréséhez, azaz a finnugor kisebbségi nyelvekre történő, felhasználói szintű nyelvtechnológiai eszközök létrehozásához így tehát szükség van az adott nyelvű szegmentálók és morfológiai elemző eszközök, továbbá szótári adatbázisok létrehozására.

A projektben elsődlegesen az udmurt és manysi nyelvekre összpontosítunk.A jelenlegi szakaszban az udmurt és manysi nyelvű korpuszok létrehozása zajlik, ezzel párhuzamosan az adott nyelvű digitális szótárak fejlesztése is folyamatban van. E szótárak szóanyagát a későbbiekben nyelvtani (morfológiai) információval is ellátjuk, így a szótárként való hasznosítás mellett a morfológiai elemzők alapjául is szolgálhatnak, melyek létrehozása szintén elkezdődött. A későbbiekben a szótári adatbázisra, korpuszainkra és a morfológiai elemzőkre építve különféle nyelvtechnológiai alkalmazásokat, például interneten szabadon elérhető szótárakat és nyelvoktató játékokat, illetve helyesírás-ellenőrzőt szeretnénk létrehozni, figyelembe véve a lehetséges jövőbeli felhasználók igényeit is.

2. A FinUgRevita projekt

A projekt célja, hogy veszélyeztetett oroszországi finnugor nyelvek beszélőit támogassuk számítógépes nyelvi eszközökkel, amelyek a felhasználók/beszélők kisebbségi nyelvi nyelvhasználatát segítik a digitális térben, valamint hogy szociolingvisztikai eszközökkel lemérjük ezen számítógépes nyelvi eszközök sikerességét. Kutatásunkkal annak a kérdésnek a praktikus megválaszolásához kívánunk hozzájárulni, hogy mivel lehet aktívan támogatni a veszélyeztetett finnugor kisebbségi nyelveket, megerősíteni a beszélő közösségeket és ilyen módon szolgálni a nyelvi revitalizációt.

A projekt nyelvtechnológiai komponenseként fel kívánjuk használni a veszélyeztetett kisebbségi finnugor nyelveken már létező nyelvi forrásokat (szótárakat és morfológiákat), hogy azokat felhasználva számítógépes eszközöket (tanulást és szövegalkotást segítő eszközöket) hozzunk létre. Ezen eszközök lehetővé teszik majd, hogy a beszélők modernizált populáris beszédmódokban használják anyanyelvüket. Úgy gondoljuk, hogy ezek az eszközök pozitív hatással lesznek a beszélők írott nyelvi tudására, anyanyelvükhöz kapcsolódó nyelvi attitűdjeikre, és végső soron a revitalizációs folyamatot segítik elő.

A projekt szociolingvisztikai komponenseként fel készülünk mérni a kifejlesztett és használatra bocsátott számítógépes eszközök sikerességét. A szociolingvisztikai méréseket az eszközök kifejlesztése és használatra bocsátása előtt és után is elvégezzük az eredmények összehasonlíthatósága érdekében.

3. Nyelvek

Az udmurt – régebbi elnevezés szerint: votják – az uráli nyelvcsaládba tartozó, kevéssé veszélyeztetett őshonos nyelv. Udmurtok nagyobb számban élnek Kazahsztánban, és szórványban Oroszország számos városában, kerületében. A legfrissebb, 2010-es népszámlálási adatok szerint az udmurtok lélekszáma 552 299 fő, az udmurt nyelvet saját bevallása szerint 324 338 fő beszéli. (Mindkét szám népszámlálásról népszámlálásra csökken.)

A manysi – régebbi elnevezés szerint: vogul – egy az uráli nyelvcsaládba tartozó, erősen veszélyeztetett őshonos nyelv. A manysi nyelvet elsősorban Nyugat-Szibériában, a Hanti-Manysi Autonóm Körzet területén beszélik. A legfrissebb, 2010-es népszámlálási adatok szerint a manysik lélekszáma 12 269 fő, a manysi nyelvet saját bevallása szerint 938 fő beszéli. (Előbbi szám népszámlálásról népszámlálásra nő, utóbbi szám folyamatosan csökken.)

A manysi és az udmurt nyelv elsősorban a családi, baráti érintkezések során használatos, nem hivatalos nyelv, nem rendelkezik gazdasági jelentőséggel, nem játszik szerepet a törvényhozásban és a politikában sem. Ugyanakkor jelen van a sajtó mellett a manysi nyelv a médiában, a kulturális életben, az interneten és az oktatásban is.

4. A FinUgRevita projekt számítógépes vonatkozásai

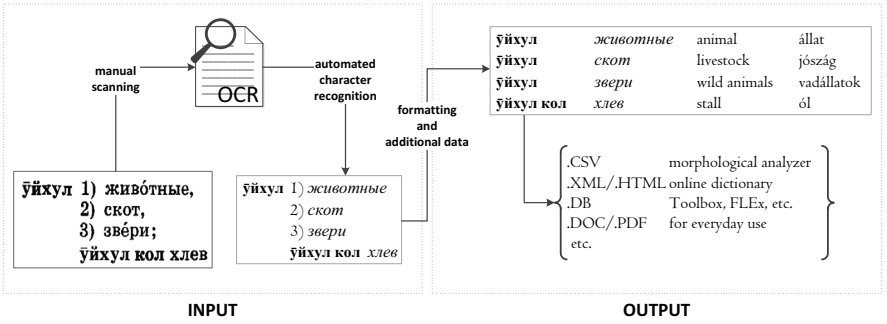
4.1. Online szótárak létrehozása

A projekt két nyelvére, az udmurtra és a manysira jelenleg folyamatban van online elektronikus szótárak készítése.

Az udmurt elektronikus szótárhoz a Kozmács István által létrehozott és szerkesztett Udmurt-magyar szótárból indulunk ki [2]. Egészen pontosan az elektronikus szótárt ez utóbbinak elektronikus verziójából (Microsoft Word dokumentum) alakítjuk át, félig automatikus módon: a dokumentum formázása alapján strukturált (CSV) formátumra konvertáljuk az anyagot, majd ezt kézzel ellenőrizve javítjuk. Az automatikus konverzió megtörtént, jelenleg a kézi javítás zajlik. A szótár kb. 13000 címszót tartalmaz.

A projekt online manysi szótára nagyobbrészt a már létező manysi–orosz és orosz–manysi szótárak [3,4,5,6] alapján készül. Az online manysi szótár tartalmazza Rombandeeva-Kuzakova 4000 szócikkes kétirányú [3] és Rombandeeva 11000 szócikkes orosz–manysi szárának [4] teljes anyagát, valamint Balandin-Vakhrusheva manysi–orosz szótárának [6] egyes válogatott szócikkeit. Ezen felül a lexikon anyagát törekszünk bővíteni a jelenkori nyelvhasználat neologizmusaival (mint például a városi környezet, életmód szavaival, az olajbányászat, jogi és közigazgatási terminusaival), melyek főként a manysi sajtóban, a *Luima Seripos*ban jelennek és honosodnak meg.

Az online manysi szótár megközelítőleg 10000 elemet tartalmaz majd. A manysi lexémákhoz angol, orosz és magyar fordítást, szófaji címkét, valamint szükség esetén ragozási paradigmát is rendelünk. A manysi lexémákat orosz fordításukkal a szótárak PDF-formátumából nyerjük ki OCR-karakterfelismeréssel. Az angol és orosz fordításokat nyelvészeink biztosítják. A 1. ábra bemutatja, hogyan készül a szótár: az OCR-t kézi javítás, kézi fordítás követi, aminek eredménye egy kereshető, digitális szótár [7].



1. ábra. A szótárépítés folyamata

4.2. Morfológiai elemzők fejlesztése

A projekt egyik legfontosabb feladata a morfológiai elemzők készítése. Első lépésben feltérképeztük a finnugor nyelvekre kifejlesztett morfológiai elemzőket és felmértük hozzáférhetőségüket.

A manysi esetében már létezik egy morfológiai elemző [8], melyet a Morpho-Logic Kft.[1] fejlesztett ki. Ennek az elemzőnek az alkalmazási és felhasználási köre viszont több szempontból eltért a FinUgRevita irányvonalaitól. Az első szempont a latin betűs átiratú folklórgyűjtések helyett napjaink cirill ortográfiájú szövegeinek elemzése (lásd Korpuszépítés 4.3 bekezdést). Másodsorban az elemző lexikona Munkácsi *Wogulisches Wörterbuch*-jára [5] épül, és Kálmán két könyvének, a *Chrestomathia Vogulica* [9] és *Wogulische Texte* [10] szövegeire lett optimalizálva. Munkácsi szótára 19. század végi szövegeken alapul, és a Kálmán köteteiben fellelhető szövegeket a 20. század első felében gyűjtötték, így az elemző lexikonából hiányzik a jelenkori 20. és 21. századi szókincs, mely az újabb szövegek elemzéséhez elengedhetetlen. Végezetül, a már létező morfológiai elemző nem szabad felhasználású. Mindezen szempontok figyelembe vételével döntöttünk egy teljesen új morfológiai elemző létrehozásáról, melyhez az Online szótárak létrehozása 4.1 bekezdésben említettek szolgálnak alapul. A szótárak feldolgozásának pillanatnyi állásában az épülő lexikon egyes szavai, szócikkei már rendelkeznek a jelentésen túl morfológiai kategorizációs jegyekkel, melyekkel a szócikkeket különböző ige- és névszóragozási paradigmákba sorolhatjuk. Az elemzés nyelvtani alapjául szolgál több manysi leíró nyelvtani munka [11,12], és anyanyelvi adatközlők segítségére is számítunk.

Az udmurt nyelv esetében a már létező udmurt elemző fejlesztőivel felvettük a kapcsolatot, velük együttműködve átalakítjuk és tovább bővítjük az általuk létrehozott elemző[2] mögött álló szótári adatbázist és nyelvtani szabályokat. Hosszabb távú terveink között szerepel az előző fejezetben is említett, készülőben lévő udmurt elektronikus szótár anyagának integrálása a már elindított online morfológiai elemzőbe.

4.3. Korpuszépítés

A projektbeli munkálatokat segítendő manysi és udmurt nyelvű korpuszokat hozunk létre. A korpuszba elsődlegesen újságcikkeket, szépirodalmi anyagokat építünk be, de más típusú szövegek felvételét is tervezzük. Jelenleg a nyers szövegek beszerzése zajlik, később ezek egységes formátumra hozatala, illetve annotációja fog megtörténni.

Az 1. táblázat összefoglalja az udmurt korpusz szavainak és karaktereinek számát diskurzustípusonként. Mint látható, a legjobban képviselt szövegtípus az újságcikkek, amelyeket az udmurt nyelvű újság, az *Udmurt Dunne*[3] biztosít számunkra, de ezen kívül található még itt gyermekek számára írt folyóiratok is,

[1] `http://www.morphologic.hu/urali/index.php?lang=hungarian&a_lang=chv`
[2] `http://giellatekno.uit.no/cgi/d-udm.eng.html`
[3] `http://udmdunne.ru/`

pl. *Kizili* és *Zechbur*. A cikkek témája elég változatos: található köztük interjú, sport- vagy kulturális hírek is.

Az internetről is töltöttünk le anyagokat, pl. a Wikipédia oldalairól és udmurt nyelvű blogokról. A legtöbb anyag már digitalizálva volt, ami megkönnyítette azok feldolgozását. A korpusz most körülbelül 70,000 tokent tartalmaz.

1. táblázat. Az udmurt korpusz diskurzustípusonkénti karakter- és szószáma

Diskurzustípus	Karakterszám	Szószám
Blogok	26615	3969
Wikipédia	32110	4293
Szépirodalom	142272	20899
Sajtó	216740	30664
Oktatás	49294	6897
Esszék	25388	3255

A manysi korpusz magját a manysi nyelven kiadott *Luima Seripos* szolgáltatja, amely 1989 óta jelenik meg. A *Luima Seripos* online archívuma[4] 46 példányt tartalmaz. Ezek a kiadások, valamint a korábbi megjelenések, együttesen 5200 cikket tartalmaznak, a korpusz mérete így több mint egymillió token.

Mivel ez a manysi újság az egyetlen stabil forrása a manysi szövegeknek, ezért ennek van legnagyobb hatása a nyelv használóira is. Abból adódóan, hogy a *Luima Seripos* képezi korpuszunk alapját, a preferált manysi ortográfiát is megválasztjuk. A 19. század végén és 20. század első évtizedeiben a manysi nyelv különböző latin betűs lejegyzései maradtak fenn nyelvészek gyűjtéseiből. A manysi írásbeliség első éveiben ugyancsak a latin betűs lejegyzést használták, bár nem központosított módon, így mindenki saját intuíciói alapján írt. A Szovjetunió nyelvi politikájának következtében 1937-től cirill alapú ortográfiára kellett váltania az itt honos kis népeknek, így a manysiul íróknak is. Ekkor ugyancsak nem alakult ki központosított változat, így a manysi helyesírás jószerével egyéni és kisebb alkotócsoportok konvencióin alapul mindmáig. Ilyen meghatározó csoport a *Luima Seripos* mindenkori szerkesztői köre és tudományos intézetek, iskolák. 1937 óta több manysi ortográfia is kialakult, kezdve a manysi morfofonológiájához nem igazodó csak orosz karaktereket alkalmazó helyesírástól, eljutva egészen napjaink ortográfiájáig, mely speciális karaktereket is szerepeltet a csak a manysiban fellelhető fonémák jelölésére (így a magánhangzóhosszúságot jelölő macron például a *а̄* „folyó" szóban, vagy *ӈ* a veláris nazális mássalhangzó [ŋ] fonéma jelölésére). A magánhangzók hosszúságának jelentésmegkülönböztető szerepe van (*ос* „felület" és *ӧс* „bárány"), ennek ellenére a jelölése csak a 20. század vége felé válik általánossá. A legutolsó ortográfiai változtatás a 2000-es évek folyamán vált konvencionálisan elfogadottá, melynek során a palatalizált zöngétlen szibiláns [s^j] jelölésére a korábbi *с* + {lágy magánhangzó: *е, ё, и, ю, я*; lágyjel: *ь*} helyett a legújabb kiadványokban a *щ* betű

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

szolgál a fonéma jelölésére. Ezen szabályokat követve a projektben is a legutolsó elfogadott ortográfiát használjuk alapértelmezettként, de a készülő morfológiai elemző a maitól eltérő transzliterációjú szövegek elemzésére is képes lesz.

5. Összegzés

Cikkünkben bemutattuk a FinUgRevita projektet, melynek célja nyelvtechnológiai eszközök létrehozása két oroszországi kisebbségi státuszú finnugor nyelvnek, az udmurtnak és manysinak. A projekt jelenlegi fázisában a két nyelv elektronikus szótárának felépítése és bővítése, valamint a világhálón fellelhető irodalmi szövegek, újságcikkek és a közösségi médiában létrejött blogok, bejegyzések nyelvi anyagából nagy tokenszámú korpuszok építése folyik. Az általunk szerkesztett szótárak lexémáin alapuló új morfológiai elemzőt fejlesztünk a korpuszok feldolgozására.

A jövőbeni terveinket tekintve elsőképpen szeretnénk a készülő szótárakat és morfológiai elemzőinket szabadon hozzáférhetővé tenni az udmurt és manysi nyelv beszélőinek, valamint a nyelveket tanulni és kutatni vágyóknak. További célunk a korpuszok morfológiai és lehetőség szerinti szintaktikai annotálása, mely alapul szolgálhat egy statisztikai szófaji egyértelműsítő és szintaktikai elemző létrehozásához.

Végül célunk olyan online nyelvi játékok tervezése és megalkotása, melyek segíthetik a nyelvtanulás folyamatát. Reményeink szerint a FinUgRevita projekt által elért eredmények szerepet játszanak majd az udmurt és manysi nyelv revitalizálásában, és a kifejlesztett nyelvtechnológiai eszközök segítséget nyújtanak, hogy az általunk támogatni szándékozott nyelvek meghonosodjanak a digitális térben is.

Köszönetnyilvánítás

A kutatás a Számítógépes eszközök a veszélyeztetett finnugor nyelvek nyelvi revitalizációjáért (FinUgRevita) nevű, FNN 107883 azonosítószámú projekt keretében valósult meg, az OTKA támogatásával.

Hivatkozások

1. Kornai, A.: Digital language death. PLoS ONE **8**(10) (2013) e77056
2. Kozmács, I.: Udmurt-magyar szótár. Savaria University Press (2002)
3. Rombandeeva, E.I., Kuzakova, E.A.: Slovar' mansijsko-russkij i russko-mansijskij. Prosvešenie, Leningrad (1982)
4. Rombandeeva, E.I.: Russko-mansijskij slovar'. Mirall, Sankt-Peterburg (2005)
5. Munkácsi, B., Kálmán, B.: Wogulisches Wörterbuch. Akadémiai Kiadó, Budapest (1986)
6. Balandin, A.N., Vahruševa, M.I.: Mansijsko-russkij slovar' s leksičeskimi paralellami iz južno-mansijskogo (kondinskogo) dialekta. Prosvešenie, Leningrad (1958)

7. Thieberger, N., Berez, A.L.: Linguistic data management. In Thieberger, N., ed.: The Oxford Handbook of Linguistic Fieldwork. Oxford University Press, Oxford (2012) 90–118
8. Prószéky, G.: Endangered uralic languages and language technologies. In: Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria (2011) 1–2
9. Kálmán, B.: Chrestomathia Vogulica. Tankönyvkiadó, Budapest (1963)
10. Kálmán, B.: Wogulische Texte mit einem Glossar. Akadémiai Kiadó, Budapest (1976)
11. Riese, T.: Vogul. Number 158 in Languages of the World/Materials. Lincom Europa, München - New Castle (2001)
12. Rombandeeva, E.I.: Mansijskij (vogul'skij) jazyk. Nauka, Moskva (1973)