Négy hatás alatt álló nyelv – Korpuszépítés kis uráli nyelvekre

Simon Eszter

MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr u. 33., e-mail: simon.eszter@nytud.mta.hu

Kivonat Cikkünkben bemutatunk egy pilot projektet, amely azt tűzte ki célul, hogy annotált nyelvi adatbázist épít négy oroszországi kisebbségi uráli nyelvre, melyek az udmurt, a tundrai nyenyec, valamint a szinjai és a szurguti hanti. A célkitűzést többek közt az indokolja, hogy az uralisztika területén inkább eklektikus adathalmazokkal találkozik a kutató, mintsem szisztematikusan annotált adatbázisokkal. Meggyőződésünk, hogy a számítógépes nyelvészet eszköztára jól használható az ilyen speciális nyelvekre történő korpuszépítés során is, és nagyban segíti az uralisták és az elméleti nyelvészek munkáját.

Kulcsszavak: korpuszépítés, számítógépes nyelvészet, uráli nyelvek, veszélyeztetett nyelvek

1. Bevezetés

Az uralisztikai kutatások jellemzően az alábbi séma szerint zajlanak. A kutató terepmunkára megy valahova Oroszországba, hazatér egy adag audió- és/vagy videófájllal, amit később feldolgoz a saját elképzeléseinek és céljainak megfelelően. Az adathalmazon kikutatott eredményeket publikálja, de az adathalmazt nem teszi publikusan hozzáférhetővé. Ha valaki valahogy mégis hozzá tud jutni az adatokhoz, akkor azzal szembesül, hogy a kutató a beszélt nyelvi anyagot valami saját lejegyzési rendszer alapján jegyezte le, amit rajta kívül senki nem használ, és nem is ismer. Dokumentáció, ami alapján meg lehetne fejteni a kódot, általában nincs, ha mégis van, akkor nincs publikálva, ha mégis, akkor nem angolul. A lejegyzés szerencsés esetben egy általánosan használt, szabadon elérhető eszközzel történik, de sokszor inkább különféle szövegszerkesztőkben, különféle házilag készített fontkészletekkel összeeszkábált, a strukturáltságnak látszatát kelteni sem igyekvő dokumentumok születnek. Ehhez jön hozzá, hogy a felvételek jogi háttere sokszor nem tisztázott, így a felhasználási lehetőségük is eléggé korlátozott.

Az elmúlt néhány évben/évtizedben a fentiekben vázolthoz képest pozitív változások zajlanak általában véve a nyelvi dokumentáció terén, szűkebben pedig az uralisztikában is. Egyre többen törekszenek arra, hogy szabadon hozzáférhetővé tegyék az adataikat, hogy sztenderd eszközöket használjanak, és hogy

valamilyen formában alkalmazzák a számítógépes nyelvészet eszközeit és/vagy módszereit ahhoz, hogy ne egy eklektikus adathalmazt, hanem egy strukturált adatbázist kapjanak eredményül.

Cikkünkben egy olyan projektet mutatunk be, amely szintén ezt a célt tűzte ki, vagyis egy nyelvi annotációt tartalmazó, sztenderd eszközökkel feldolgozott és sztenderd formában, szabadon elérhető strukturált adatbázis létrehozását oroszországi kisebbségi uráli nyelvekre.

A projekt címe: *Az uráli nyelvek mondattanának változása aszimmetrikus kontaktushelyzetben*, időtartama másfél év (2016. február – 2017. július), befogadó intézménye az MTA Nyelvtudományi Intézete, projektvezetője É. Kiss Katalin. A projektet az NKFI támogatja, azonosítója: ERC_HU_15 118079. Ez egy pre-ERC projekt, amelynek az a célja, hogy lehetőséget adjon egy jövőbeli ERC[1] pályázat elméleti és módszertani alapjainak lefektetésére. A cikkben ismertetett elméleti és módszertani megfontolások a folyamatban levő pilot projekt során lettek kidolgozva, de természetesen a majdani ERC projektre is vonatkoznak.

A projektnek két fő célja van. Az elméleti cél egyrészt a kihalás szélén álló rokon nyelvek sajátos mondattani tulajdonságainak a leírása, másrészt ezen nyelvek szintaktikai változásainak vizsgálata, amelyek feltételezésünk szerint az orosz nyelv erőteljes hatására mennek végbe. A projekt másik célja egy annotált korpusz létrehozása *udmurt*, *tundrai nyenyec*, *szinjai* és *szurguti hanti* nyelvű, írott és beszélt nyelvi szövegekből, amely lehetővé teszi az uráli–orosz kontaktushatás kutatását. Ahhoz, hogy változásokat tudjunk detektálni, különböző korokból származó szövegeket kell gyűjtenünk és összehasonlítanunk. Az Oroszország területén beszélt kisebbségi uráli nyelvek esetében a legrégebbi írott nyelvi források a 19. század végéről származnak, amikor szervezett expedíciók keretében indultak terepmunkára etnográfusok, nyelvészek és egyéb szakemberek, hogy feltérképezzék a rokon nyelveket. Vagyis az általunk vizsgált régi szövegek a 19. század végéről – 20. század elejéről származnak. Emellett mai anyagot is gyűjtünk, nyomtatott és elektronikus forrásokból, illetve terepmunkán gyűjtött beszélt nyelvi adatokból.

A pilot projekt keretein belül mindegyik nyelvnek mindkét korából származó szövegeket gyűjtünk, és állítjuk elő legalább az eredeti szöveg kitisztított változatát. Az adatok minden szintű feldolgozását, IPA-átiratát, teljes morfológiai elemzését és legalább angol fordítását viszont csak kb. 4000 token/kor/nyelv mennyiségű adatra tervezzük a pilot projektben. Természetesen a majdani ERC projektben ennek sokszorosára lesz szükség ahhoz, hogy az egyes nyelvi jelenségek változásáról tényleges következtetéseket lehessen levonni.

A cikk további része az alábbiak szerint épül fel. A 2. fejezet a korpuszépítés gyakorlati lépései mögött meghúzódó elméleti és módszertani megfontolásokat mutatja be. A 3. fejezet ismerteti, hogy milyen szövegeket gyűjtöttünk és honnan, majd a 4. fejezet bemutatja az egyes szövegfeldolgozó lépéseket. Az 5. fejezet a korpusz felépítését írja le, és végül a 6. fejezet tartalmazza a konklúzióinkat és a jövőbeli terveinket.

[1] https://erc.europa.eu/

2. Elméleti megfontolások

A projekt nyelvei mind veszélyeztetettek és hiányosan dokumentáltak, de azért mutatkozik köztük némi különbség. Az udmurt nyelv több szempontból is kilóg a többi közül. Egyrészt Udmurtia egyik hivatalos nyelve, másrészt a nyelvi veszélyeztetettséget jelölő EGIDS-skálán [9,3] az 5., vagyis az *írott* kategóriába tartozik. Ez utóbbi annyit tesz, hogy a nyelvet napi szinten használják, és létezik egy sztenderd irodalmi változata, de az nem annyira terjedt el.

A projekt másik három nyelve mind szibériai nyelv, és mind a 6b, vagyis *veszélyeztetett* kategóriába tartoznak az EGIDS-skálán. Ezeket a nyelveket manapság már szinte csak az idősebb generáció használja, ők is csak családi és informális körben. Nem hivatalos nyelvek, továbbá alacsony presztízsűek, és a rájuk irányuló revitalizációs törekvések sem mondhatóak nagy számúnak és sikeresnek.

Ezek a tényezők több olyan következménnyel járnak, amelyeket figyelembe kell venni a korpuszépítés során, és amelyek a jól dokumentált, sok beszélős nyelvek esetében nem feltétlenül játszanak fontos szerepet.

A korpuszépítés során figyelembe vett egyik fő kritérium az volt, hogy – lehetőségeinkhez mérten – kövessük a nyelvi dokumentáció alapelveit. A nyelvi dokumentáció egy nyelv adatainak rögzítését, annotálását, megőrzését és disszeminációját jelenti, azaz gyűjtést, feldolgozást, annotációt, közzétételt, archiválást és tárolást [20]. Projektünkben a himmelmanni [6] értelemben vett elsődleges adatokat rögzítjük és dolgozzuk fel. Ezek olyan kommunikációs eseményekből származó nyelvi adatok, amelyek a hétköznapi nyelvhasználatot tükrözik, például dialógusok, elbeszélések, élettörténetek, vagyis nem irányított beszélgetések és nem feldolgozott szövegek, szólisták, kérdőívek.

A nyelvi dokumentáció súlypontjai az elmúlt évtizedekben áthelyeződtek (vö. [1,17]). A nyelvi dokumentáció új szemléletet és új eszközöket használ, a leírásban teljességre, egységességre és összehasonlíthatóságra törekszik. Ez utóbbiakra törekszünk mi is a korpuszépítés során, amelyek betartásához a számítógépes nyelvészeti eszközök és módszerek használata segítséget nyújt.

A teljességre törekvés azt jelenti, hogy abban a szellemben kell gyűjteni az anyagot, hogy az minél szélesebb körben használható legyen majd. Ezért az adatbázis-építés során arra törekszünk, hogy a lehető legtöbb szerzőtől válasszunk szöveget, és ezek minél több társadalmi osztályt, kort, nemet, dialektust és műfajt öleljenek fel. Továbbá az is fontos, hogy az eredeti felvétel, vagyis az audió- és/vagy videóanyag is elérhető legyen, hogy a leírások és következtetések ellenőrizhetők legyenek. Ahhoz, hogy az adatbázis tényleg hasznosítható legyen más területeken, így például szociolingvisztikai és antropológiai kutatásokhoz is, gazdagon kell metaadatolni minden nyelvi adatot.

Az egységesség és összehasonlíthatóság az adatbázis-építés minden szintjén megjelenik. Fontos egyrészt, hogy a nyelvi annotáció során nem követünk semmilyen nyelvészeti paradigmát, másrészt viszont szigorúan követünk bizonyos nemzetközi sztenderdeket, hogy a nyelvek és az eszközök közötti átjárhatóságot biztosítsuk.

A különböző nyelvű, különböző ábécét használó, különböző lejegyzést követő szövegek egységes reprezentációjához sztenderd Unicode-karaktereket használunk a teljes korpuszban (a projektben használt lejegyzési, átírási és írásrendszerekről részletesebben lásd a 4.1. fejezetet).

A hangok szintjén a Nemzetközi Fonetikai Ábécét (International Phonetic Alphabet, IPA) követjük. Erre azért van szükség, mert az uráli nyelvek lejegyzői hagyományosan a Setälä-féle [15] átírási rendszert használják (részletesebben lásd a 4.2. fejezetet), amely egyrészt nem egy egységes rendszer, másrészt nem ismert az uralisztikán kívül, ezért minden szövegnek automatikusan legeneráljuk az IPA-átiratát is.

A morfológia szintjén a lipcsei glosszázási szabályokat (Leipzig Glossing Rules, LGR)[2] követjük. A tokenek és a hozzájuk tartozó morfológiai információk egymásnak megfeleltetve, párhuzamosítva vannak megjelenítve. A glosszák az említett nyelvekre elérhető morfológiai elemzők kimenetéből állnak elő automatikus konvertálással (további részletekért lásd a 4.3. fejezetet), amiből az következik, hogy a morfológiai annotáció csak akkor lesz morféma szinten is megfeleltetve, ha az elemző képes szegmentálásra. Ebben az esetben, az LGR szabályait követve, kötőjellel választjuk el egymástól a morfémákat, illetve az őket jelölő kódokat. Az LGR tartalmaz egy ajánlott címkelistát is, amelyet követünk, de némileg kiegészítve, tekintve, hogy az eredeti lista nem fedi le az általunk elemzett nyelvek minden morfológiai jelenségét.

A nemzetközi szabványok követése az általunk alkalmazott formátumok terén is jelentkezik, ami jelen nyelvek esetében azért is fontos, mert minden nyelvi dokumentációs és nyelvfeldolgozó eszköz, amely ezekre elérhető, különböző ki- és bemeneti formalizmusokat követ, amelyek között a szabványos formátumok biztosítják az átjárhatóságot. Az általunk előállított összes szöveges állomány UTF-8 karakterkódolású sima szöveg fájl. A tokenszintű annotációk oszlopok formájában vannak reprezentálva sztenderd `tsv` fájlokban, amelyek bemenetül szolgálhatnak további nyelvfeldolgozó eszközök számára, vagy könnyen átalakíthatók `XML`-fájlokká.

3. Szöveggyűjtés

Ahogy fentebb említettük, arra törekszünk, hogy a korpusz reprezentatív mintája legyen az adott nyelvi közösség nyelvhasználatának. Ezt a törekvésünket azonban a 2. fejezetben kifejtett tényezők nagyban befolyásolják. Mivel a projektben vizsgált szibériai nyelvek esetében nemigen beszélhetünk sztenderd írásbeliségről, továbbá a nyelvet elsősorban az idősebb generáció használja, akik nem termelnek napi szinten elektronikus szöveges adatot, ezen nyelvek esetében nem támaszkodhatunk olyan, viszonylag könnyen elérhető forrásokra, mint a blogok, tweetek vagy a napi sajtó. Az is nehezíti továbbá a szövegek begyűjtését, hogy a korábbi, terepen gyűjtött anyagokat a kutatók jellemzően nem teszik publikussá. Ha mégis elérhető elektronikus formában valamilyen anyag, akkor az

[2] https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

egyrészt nem túl sok, másrészt inkább eklektikus adathalmaz, mint szisztematikusan annotált korpusz.

Mindezekből következik, hogy a szöveggyűjtésnél eléggé meg van kötve a kezünk. A régi szövegek közé olyan folklór szövegeket válogattunk, amelyeket a 19. század végén – 20. század elején gyűjtöttek, és maga a terepen járt kutató adta közre a maga lejegyzési szisztémája alapján. A régi szinjai hanti szövegek Wolfgang Steinitz [16] gyűjtéséből származnak az 1930-as évekből, míg a szurguti hanti szövegeket Heikki Paasonen [18] gyűjtötte 1900-01-ben a Jugán folyó környékén. A régi udmurt szövegek két forrásból származnak: egyrészt Yrjö Wichmann [19] gyűjtéséből, ami 1901-ben lett publikálva, másrészt Munkácsi Bernát 1887-es terepmunkájából [10]. A régi tundrai nyenyec szövegek forrása Toivo Lehtisalo 1911–12-es gyűjtése [8]. Annak ellenére, hogy ezek mind folklór szövegek, vagyis ugyanabba a műfajba tartoznak, a szövegválogatást igyekeztünk úgy végezni, hogy a dialektusok és az adatközlők kora és neme szerint kiegyensúlyozott legyen. Az összes elérhető metaadatot összegyűjtjük, és táblázatba rendezve közreadjuk a projekt weboldalán.

Az új szövegek sokkal inkább különböző műfajú forrásokból származnak: az új hanti adatok lejegyzett interjúkat tartalmaznak, míg az udmurt szövegek a *Mynam malpanësy*[3] és a *Marajko*[4] nevű blogokból származnak. A modern tundrai nyenyec adat tartalmaz újságcikkeket a *Njar'jana Ngerm* című újságból, valamint új gyűjtésű folklór szövegeket Labanauskas [7] és Puškarëva–Chomič [14] gyűjtéseiből.

A beszélt nyelvi adatok a projektrésztvevők terepmunkái során gyűjtött és a jövőben gyűjtendő anyagaiból áll össze. Ezek a felvételek az ELAN-ban[5] lesznek lejegyezve és illesztve. Terveink szerint az új szövegek ugyanabból a régióból lesznek gyűjtve, ahonnan a régiek is származnak, hogy a nyelvjárási különbségeket kiküszöböljük a szintaktikai változások vizsgálata során.

4. Szövegfeldolgozás

A korpuszépítési workflow első lépése az eredeti szöveges anyag előállítása és egységes formátumra hozása, ezt írja le a 4.1. fejezet. A különféle lejegyzési és átírási rendszerek közötti átjárást biztosítanunk kell; az ehhez szükséges konverziós lépésekről a 4.2. fejezet tudósít. A korpusz morfológiai annotációt is tartalmaz, amelynek leírása a 4.3. fejezetben található.

4.1. Az eredeti szöveg előállítása

A *beszélt* nyelvi adatok feldolgozásának első lépése a lejegyzés, más néven transzkripció. Az uralisztikában a FUT (Finno-Ugric transcription) vagy más néven uráli fonetikai ábécé az elterjedt, amelyet Eemil Nestor Setälä [15] publikált 1901-ben azzal a szándékkal, hogy az uralisták által használt lejegyzési rendszereket

[3] http://udmurto4ka.blogspot.hu/
[4] http://marjamoll.blogspot.hu/
[5] http://tla.mpi.nl/tools/tla-tools/elan/

egységesítse. Ennek ellenére a FUT-ba sorolt lejegyzések nem alkotnak egy következetes rendszert, sőt igen jellemző, hogy ugyanannak a hangnak a jelölésére más és más karaktert használnak.

Miután megtörtént a beszélt nyelvváltozat lejegyzése, az adat onnantól kezdve ugyanazokon a feldolgozási lépéseken megy keresztül, mint az írott nyelvi anyag. A régen lejegyzett és kiadott szövegek is lejegyzett beszélt nyelvi anyagnak számítanak a további feldolgozás szempontjából.

Az általunk feldolgozni kívánt *írott* nyelvi adatok egy része csak nyomtatott könyv formájában volt elérhető, ezért ezeket beszkenneltük, majd optikai karakterfelismerő (OCR) program segítségével jutottunk hozzá a szöveghez. A korpuszunkban található nagyszámú lejegyzési és írásrendszer kezelése miatt az OCR programmal szemben alapvető elvárásunk volt a taníthatóság. Az Abbyy FineReader Professional Edition[6] mellett döntöttünk, ami ugyan nem nyílt forráskódú, de meglehetősen könnyen tanítható, és elég jó minőségű kimenetet ad.

Bizonyos dokumentumokat a webről töltöttünk le; ebben az esetben `HTML`-forrásokból és `PDF`-fájlokból kellett kinyernünk a szöveget. A kimenetet minden esetben kézzel ellenőriztük, hogy a következő feldolgozó lépésben minél tisztább anyaggal dolgozhassunk.

A szabványosság előnyei miatt a teljes korpuszt sztenderd UTF-8 kódolású Unicode-karakterekkel tároljuk és jelenítjük meg. Mindenképpen szükséges egy az egész korpuszra kiterjedő szigorúan egységes formátum, ez teszi lehetővé, hogy a lekérdezéseket az egész anyagra vonatkoztathassuk. Ezt csak úgy biztosíthatjuk, ha következetesen betartjuk azt az alapelvet, hogy azonos dolgokat mindig ugyanúgy, különbözőeket pedig mindig eltérően jelölünk.

Ennek eléréséhez az első lépés az volt, hogy létrehoztunk egy egységes karaktertáblát, amelyben minden nyelv minden transzkripciós, transzliterációs és írásrendszerének minden karaktere szerepel a Unicode-kódjával és -nevével, valamint Prószéky-kódjával egyetemben. Ez a kódtábla van használva minden szövegfeldolgozó lépésnél: ezekkel a karakterekkel történik a hangzó szövegek lejegyzése, ezekre a karakterekre tanítjuk be az optikai karakterfelismerőt, ezekre a karakterekre normalizáljuk a különböző forrásokból származó szövegeket, és ezek szolgáltatják a különböző irányú konverziók bemeneti és kimeneti karakterállományát is (lásd a 4.2. fejezetet).

A következő lépésben ellenőrizzük és normalizáljuk az összes szöveget egy `Perl`-szkript[7] segítségével, amely kilistázza a dokumentumban szereplő Unicode-karaktereket. A lista alapján könnyedén felismerhetők és eltávolíthatók az idegen nyelvű részek, illetve a nem helyesen használt karakterek lecserélhetők.

4.2. Átírás és konverzió

A transzkripcióval szemben meg kell különböztetnünk a transzliterációt, amely egy már írott formában létező nyelvi adat átírása egy másik írás- vagy jelölési

[6] http://finereader.abbyy.com/
[7] https://gist.github.com/takdavid/3fa2cc3ae21aa96da24b8bd90b8c63b0

rendszerre. Ahogy említettük, az adatbázisunk tartalmaz minden szöveget legalább az eredeti lejegyzésében, amelyet a nyelv dokumentálója használ, valamint IPA-átírásban is. Ez utóbbit azért tartjuk fontosnak, mert így nem csak az uralisztika kutatói, hanem más nyelvészek is olvasni és használni tudják az anyagot. Továbbá – mivel az érintett nyelvek írásrendszere a cirill ábécén alapszik – megőrizzük az eredeti cirill írást, amennyiben van ilyen. Ha nincs, de szükség van rá a morfológiai elemzőhöz, akkor egy konverziós lépés során előállítjuk. Ugyanígy járunk el a különféle FUT-típusú lejegyzésekkel is: mivel bizonyos morfológiai elemzők csak bizonyos módon lejegyzett szövegeket fogadnak el inputként, ezeket is elő kell állítani egy konverziós lépés során. (A morfológiai elemzőkről lásd a 4.3. fejezetet.)

A projektben vizsgált négy nyelvre összesen 11 konverziós irány van, amelyekre konvertereket fejlesztettünk. A régi szinjai hanti szövegek eredetileg Steinitz lejegyzésével készültek, aki a saját FUT-jellegű rendszerét használta. Ezt konvertáljuk először IPA-ra, aztán arra a szintén FUT-jellegű ábécére, amelyet az általunk használt morfológiai elemző fejlesztői alkalmaztak. Az új szinjai hanti szövegek lejegyzése már eleve ez utóbbi szerint zajlik.

A régi szurguti hanti szövegeket az Ob-Ugric Database (OUDB)[8] fejlesztői bocsátották a rendelkezésünkre, és mivel ők csak IPA-ban tették elérhetővé az anyagukat, nekünk is csak IPA-átiratunk van. A modern szurguti hanti szövegek viszont a mai cirill betűs hanti írással íródtak, amelyet először átkonvertálunk a Csepregi Márta [2] által alkotott és a hanti nyelvet kutatók körében széles körben használt átírásra, majd ebből állítjuk elő az IPA-verziót.

Az udmurt nyelv esetében négy különböző konverterre van szükség. Először létrehoztuk a konverziós szabályokat a Munkácsi–IPA és a Wichmann–IPA irányokba, majd az IPA-verziót konvertáljuk cirill betűs írásmódra. Ez utóbbira azért van szükség, mert az udmurtra fejlesztett morfológiai elemzők mindegyike cirill betűs bemenetet vár. Az új udmurt szövegek esetében az irány fordított, vagyis a cirill szöveget konvertáljuk IPA-ra.

A régi tundrai nyenyec szövegek bizonyos értelemben kivételt képeznek. Lehtisalo olyan bonyolult transzkripciós rendszert dolgozott ki, amelyre se az IPA-átírás elkészítéséhez, se a morfológiai elemzőhöz nincs szükség, továbbá egy részük nem is lenne reprezentálható sztenderd Unicode-karakterekkel. Ezért a Lehtisalo-szövegek OCR-ezésénél egy Lehtisalo–Hajdú leképezést használtunk, így ezek a szövegek már eleve Hajdú Péter [5] transzkripciója alapján készültek el. Ez utóbbi lett IPA-ra, majd cirillre konvertálva, az utóbbi a morfológiai elemzőhöz. A modern nyenyec szövegekkel hasonló a helyzet, mint az udmurttal: a cirill betűs modern nyenyec írásnak is elkészítjük az IPA-konverzióját.

A konverzió első lépéseként az adott nyelv szakértői átírási szabályokat definiáltak. Ezek lettek kiterjesztett reguláris kifejezéseket tartalmazó helyettesítési parancsokká átalakítva, és így beadva a `sed` parancsnak segédfájlként egy `-f` kapcsolóval. Vagyis ez egy szabályalapú rendszer, annak minden tipikus előnyével és hátrányával. Hátrányai közé tartozik, hogy nyelvfüggő, sőt jelen esetben irányfüggő, vagyis nem vihető át egy másik konverziós irányra változtatás nélkül.

[8] http://www.oudb.gwi.uni-muenchen.de/

Ezen kívül, ha sok szabállyal dolgozunk, amelyeknek fontos a sorrendje is, nem mindig egyszerű fejben tartani az összeset, így könnyű hibázni, ami tökéletesen rossz eredményhez vezethet. Van viszont egy nagy előnye a szabályalapú rendszereknek, mégpedig az, hogy magas pontosságot produkálnak. Mivel az automatikusan konvertált szövegeket nyelvész szakértők ellenőrzik a projektünkben, mi a magas pontosság mellett voksoltunk, a fent említett hátrányok ellenére is.

4.3. Morfológiai elemzés

A korpusz egy része morfológiai szintű annotációt is tartalmaz. Ezekben a szövegmintákban minden tokennél megadjuk a lemmát, a szófaji címkét és az angol glosszát. Ezek az információk a rendelkezésre álló morfológiai elemzők kimeneteiből lesznek konvertálva. Ehhez első lépésben meg kell csinálni egy leképezést, amely a különböző morfológiai elemzők által használt címkekészletet képezi le az általunk létrehozott egységes morfológiai címkekészletre. Ez utóbbiban és a glosszázás során általában is az LGR konvencióit és rövidítéseit követjük, kisebb kiegészítésekkel.

Az általunk vizsgált négy nyelvből háromra létezik morfológiai elemző, amelyet tudunk használni a morfológiai annotáció előállításának nyelvtechnológiai támogatására. Ennek ellenére az annotáció nem teljesen automatikusan készül, hanem kézi javítást is igényel.

A legismertebb szövegfeldolgozó keretrendszer kis uráli nyelvek nyelvtechnológiai támogatására a Giellatekno[9], amelynek keretein belül mások mellett helyesírás-ellenőrzők, digitális szótárak és morfológiai elemzők is fejleszthetők. Ez utóbbi már létezik, bár folyamatosan fejlesztés alatt áll, az udmurt, az északi hanti és a tundrai nyenyec nyelvekre (az északi hantinak egy aldialektusa a szinjai hanti).

Emellett létezik egy másik morfológiaielemző-csomag is kis uráli nyelvekre, így udmurtra és szinjai hantira, a MorphoLogic Kft. és az MTA Nyelvtudományi Intézetének közös munkájának eredményeként [12,4]. Ezek az elemzők nem szabad forráskódúak, hanem egy online felületen keresztül érhetők el[10]. A kimenetük egy `HTML`-fájl, amely minden beadott token minden lehetséges elemzését tartalmazza. A kézi munka megkönnyítéséhez egy webes felületet használunk, amely eredetileg ómagyar szövegek morfológiai egyértelműsítéséhez lett kifejlesztve [13], de némi módosítással a mi céljainkra is használható. A felhasználó az egyértelműsítendő token fölé egerészik, majd az összes elemzést tartalmazó legördülő menüből kiválasztja a helyes elemzést. Azokhoz a szavakhoz, amelyeket nem ismert fel az elemző, kézzel kell bevinni a helyes elemzést. Ez a webes interfész a Giellatekno outputján is használható.

A szinjai hanti és az udmurt szövegek elemzésére a morphologicos elemzőt használjuk, mert ez morféma szinten szegmentált kimenetet ad, továbbá a magyar (és a szinjai hanti esetében az angol) fordítást is előállítja.

[9] http://giellatekno.uit.no/
[10] http://www.morphologic.hu/urali/

A tundrai nyenyec szövegek elemzésére a Giellatekno elemzőjét használjuk. Mivel az elemző szótára a tundrai nyenyecnek csak egy dialektusába tartozó szavakat tartalmazza, valamint a nyelvtanfájlok egy korábbi nyelvtan alapján készültek, terveink között szerepel egyrészt a szótár bővítése egyéb nyelvjárásokba tartozó elemekkel, másrészt a nyelvtanfájlok update-elése a legújabb nyelvtan [11] alapján.

Sajnos a negyedik nyelvre, a szurguti hantira nem tudunk elérhető morfológiai elemzőről, de azért megpróbáltunk erre a nyelvre is valamilyen automatikus támogatást nyújtani. Amit alkalmaztunk, az egy végtelenül egyszerű memóriaalapú megoldás. Zipf törvénye alapján tudjuk, hogy a néhány leggyakoribb szó lefedi a teljes szöveg nagy százalékát. Ebből kiindulva kilistáztuk a modern szurguti hanti szöveg minden olyan tokenjét, amely legalább ötször előfordul. Ezekhez egy nyelvész szakértő kézzel hozzárendelte a szófaji kódot, az inflexiós címkéket és a lemma angol fordítását. Ezzel a glosszák több mint 60%-át tudjuk automatikusan generálni, ami nagy mértékben csökkenti a kézi munka mennyiségét.

5. A korpusz felépítése

A korpusznak három fő annotációs szintje van. A transzkripció és a transzliteráció, vagyis az eredeti szöveg és az átírások szintje, a morfológiai elemzés szintje, valamint a fordítások szintje. Minden dokumentumhoz minden szinten legalább egy szövegverziónak meg kell lennie. Ezek a kötelező verziók sorrendben a következők: az IPA-átírás, a lemma, a szófajcímke és az angol glossza, valamint az angol fordítás. Az átírások és a morfológiai elemzés szintjén az annotáció tokenszintű, vagyis minden egyes tokenhez megadjuk legalább az IPA-átiratát és az előbb felsorolt morfológiai információkat. A fordítás ezzel szemben mondatszintű annotáció, vagyis teljes mondatokhoz rendelünk legalább angol, de sokszor magyar, német és orosz fordítást is. Ez utóbbiak teljes mértékben kézzel készülnek.

A token- és mondatszintű annotációkat tartalmazó szövegfájlokat beimportáljuk az ELAN-ba, ahol mondat szinten időben illesztve lesznek az audió- vagy videóanyaghoz. Az ELAN az annotációs szinteket horizontális szintekként jeleníti meg, amit az 1. táblázat illusztrál egy tundrai nyenyec példával.

6. Konklúzió és jövőbeli kutatási irányok

Cikkünkben bemutattunk egy pilot projektet, amely azt tűzte ki célul, hogy annotált nyelvi adatbázist épít négy oroszországi kisebbségi uráli nyelvre. A célkitűzést az indokolja, hogy ezeken a nyelveken jól vizsgálható az orosz–uráli kontaktushatás, amely a projekt elméleti célja, valamint hogy az uralisztika területén inkább eklektikus adathalmazokkal találkozik a kutató, mint szisztematikusan annotált adatbázisokkal. Meggyőződésünk, hogy a számítógépes

1. táblázat. Token és mondat szinten illesztett tundrai nyenyec szöveg.

YRK Hajdú:	jā	mī̄daxana	amkerta	jaŋkūwi
YRK IPA:	ja	mi:daxana	ămkerta	jăŋkuwi
YRK Cyrillic:	я	мыдахана	амкэрта	яӈкувы
lemma:	я	мы	ӈамгэ	яӈгось
POS:	N	Ptcp	Pron.neg	V
glossza:	earth	create.IPFV.PTCP.LOC	nothing	neg.EX.INFER.3SG
ENG:	when the earth was created, there was nothing			
GER:	zur zeit der erschaffung der erde gab es nichts			
HUN:	a Föld teremtésének idején nem volt semmi			

nyelvészet eszköztára jól használható az ilyen speciális nyelvekre történő korpuszépítés során is, és nagyban segíti az uralisták és az elméleti nyelvészek munkáját.

A cikkben leírt elméleti és módszertani megfontolások nem csak a pilot projektben, hanem a majdani ERC-projektben is hasznosíthatóak lesznek, míg a pilot projekt során épített korpusz anyaga a jövőben bővítésre szorul.

A korpuszépítés során követjük az open access filozófiáját, amelynek két vetülete is van. Egyik, hogy törekszünk arra, hogy szabadon elérhető eszközöket használjunk, valamint hogy újrahasznosítsunk már valamilyen formában publikált adatokat is. Másrészt a projekt eredményeképpen előálló minden szöveges és feldolgozó erőforrást szabadon hozzáférhetővé teszünk a projekt weboldalán: http://www.nytud.hu/oszt/elmnyelv/urali/adatbazisok.html.

Távolabbi terveink között szerepel, hogy az adatbázis ne csak letölthető formában legyen elérhető, hanem egy online lekérdező felületen keresztül is, amely a számítógépes eszközök használatában kevésbé jártas kutatók számára is lehetőséget nyújt az adatok használatára. Ezenfelül, a hosszú távú megőrzés jegyében, az általunk létrehozott összes adatot szeretnénk elérhetővé tenni egy nemzetközi nyelvi archívumon keresztül is, mint amilyen a The Language Archive[11] által működtetett DOBES (Documentation of Endangered Languages) korpusz.

7. Köszönetnyilvánítás

A projektet az NKFI támogatja, a pályázat azonosítója: ERC_HU_15 118079.

Az elméleti alapok lefektetésében és a korpuszépítésben több kutató is részt vett; a korpusz nélkülük nem jött volna létre. Ők név szerint: Asztalos Erika, Gugán Katalin, Kalivoda Ágnes, Mus Nikolett, Nguyen-Dang Nóra Lien, Ruttkay-Miklián Eszter, Tánczos Orsolya.

Külön köszönettel tartozunk az OUDB projekt vezetőjének, Elena Skribniknek, hogy rendelkezésünkre bocsátotta a Paasonen-szövegeket; Schön Zsófiának,

[11] https://tla.mpi.nl/

hogy rendkívül sokat segített a szurguti hanti szövegek IPA-átírásával kapcsolatban; A. S. Pesikovának és A. N. Volkovának, hogy engedélyezték nekünk az általuk felvett és lejegyzett interjúk felhasználását.

Hivatkozások

1. Blokland, R., Fedina, M., Gerstenberger, C., Partanen, N., Rießler, M., Wilbur, J.: Language documentation meets language technology. In: First International Workshop on Computational Linguistics for Uralic Languages. pp. 8–18. No. 2 in Septentrio Conference Series (2015)
2. Csepregi, M.: Szurguti osztják chrestomathia. Szeged (2011)
3. Fazakas, N.: Újabb fejlemények a nyelvi revitalizáció kutatásában. Nyelv- és irodalomtudományi közlemények LVIII.(2), 155–164 (2014)
4. Fejes, L., Novák, A.: Obi-ugor morfológiai elemzők és korpuszok. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010). pp. 284–291. Szegedi Tudományegyetem (2010)
5. Hajdú, P.: Chrestomathia Samoiedica. Tankönyvkiadó, Budapest (1989)
6. Himmelmann, N.P.: Linguistic data types and the interface between language documentation and description. Language Documentation and Conservation 6, 187–207 (2012)
7. Labanauskas, K.I.: Neneckij fol'klor. Mify, skazki, istoričeskie predanija. Vyl. 5. Krasnojarsk (1995)
8. Lehtisalo, T.: Juraksamojedische Volksdichtung. Suomalais-Ugrilainen Seura, Helsinki (1947)
9. Lewis, M.P., Simons, G.F.: Assessing endangerment: Expanding Fishman's GIDS. Revue Roumaine de Linguistique 55(2), 103–120 (2010)
10. Munkácsi, B.: Votják népköltészeti hagyományok. Magyar Tudományos Akadémia, Budapest (1887)
11. Nikolaeva, I.: A Grammar of Tundra Nenets. Mouton de Gruyter (2014)
12. Novák, A.: Morphological Tools for Six Small Uralic Languages. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). pp. 925–930. ELRA (2006)
13. Novák, A., Orosz, G., Wenszky, N.: Morphological annotation of Old and Middle Hungarian corpora. In: Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 43–48. Association for Computational Linguistics, Sofia, Bulgaria (August 2013), http://www.aclweb.org/anthology/W13-2706
14. Puškarëva, J.T., Chomič, L.V.: Fol'klor nencev. Novosibirsk (2001)
15. Setälä, E.N.: Über Transskription der finnisch-ugrischen Sprachen. Finnisch-ugrische Forschungen 1, 15–52 (1901)
16. Steinitz, W.: Ostjakologische Arbeiten. Akadémiai Kiadó, Budapest (1975)
17. Szeverényi, S.: Rendkívül rövid bevezetés a dokumentációs nyelvészetbe. In: Szeverényi, S., Szécsényi, T. (eds.) Érdekes nyelvészet, pp. 146–157. JATE Press, Szeged (2015)
18. Vértes, E. (ed.): Heikki Paasonens surgutostjakische Textsammlungen am Jugan. Neu transkribiert, bearbeitet, übersetzt und mit Kommentaren versehen von Edith Vértes, Mémoires de la Société Finno-Ougrienne, vol. 240. Suomalais-Ugrilainen Seura, Helsinki (2001)

19. Wichmann, Y.: Wotjakische Sprachproben II. Sprichwörter, Rätsel, Märchen, Sagen und Erzählungen. Helsinki (1901)
20. Woodbury, A.C.: Language documentation. In: Austin, Peter K.; Sallabank, J. (ed.) The Cambridge Handbook of Endangered Languages, pp. 159–186. Cambridge University Press (2011)