

Zárójelentés

Az uráli nyelvek mondattanának változása aszimmetrikus kontaktushelyzetben című,
118079. számú, ERC_HU_15 jelű OTKA-projektumról

A projektum feladatul tűzte ki (1) az oroszországi uráli nyelvekben, különösen az udmurtban, a hantiban és a nyenyecben orosz hatásra bekövetkező gyors és alapvető szintaktikai változások tanulmányozását, (2) e veszélyeztetett, illetve a kihalás szélén álló nyelvek kevésbé ismert szintaktikai sajátosságainak elemzését; (3) a konkrét nyelvi változásokkal, illetve az uráli nyelvek sajátosságaival kapcsolatos elemzéseink elméleti tanulságainak megfogalmazását, (4) korai és jelenkori szövegekből álló udmurt, hanti és nyenyec adatbázis létrehozását.

E feladatok mindegyikében jelentős eredményeket értünk el – amint ezt projektumunk honlapja (<http://www.nytud.hu/oszt/elmnyelv/urali/index.html>) tanúsítja. A kutatás eredményei az alábbiakban foglalhatók össze.

1. Orosz hatásra bekövetkező mondattani változások

1.1. Változások az alapszórendben

Tanulmányoztuk, hogyan terjed az udmurtban, a hantiban és a nyenyecben az uráli SOV szórend rovására az orosz mintájú SVO szórend. Hasonló szórendi változások sok (például indo-európai) nyelvben rekonstruálhatók, de évszázadokkal korábban lezajlott változásokról lévén szó, lefolyásukkal, közbülső állomásaikkal kapcsolatban csak hipotézisek fogalmazódhattak meg. Az SOV→SVO változás kiváltója egyes nézetek szerint az ige (V) előrevivése egy T(emporális) fejpozícióba (Kiparsky 1996); mások szerint a tárgy hátravivése (jobboldali diszlokációja/ topikalizációja) és alaphelyzetű bővítményként való újraértelmezése (van Kemenade & Milicev 2012; Taylor and Pintzuk 2012). Mi az utóbbi nézet mellett találtunk bizonyítékokat. Mindhárom vizsgált nyelv esetében összehasonlítottuk az adott nyelv régi és új szövegeinek szórendi sémáit. Azt találtuk, hogy régi (20. század eleji) szövegekben a nem-igevégű mondatok aránya igen alacsony, és posztverbális helyzetben jellemzően a szövegelőzményből már ismert információt kifejező összetevő, jobbra kihelyezett topik áll. Később a régi információt kifejező összetevőkön túl új információt hordozó fókusz is kerülhet az ige mögé. Végül információszerkezet szempontjából semleges, alapszórendű mondatokban is megjelenik a VO szórend. Tehát a VO szórend terjedése, általánosabbá válása a jobbra kihelyezett összetevők információszerkezeti funkciókörének bővülésével, majd a posztverbális elemek alapszórendüként való átértelmezésével történik. Mivel az oroszban az információs fókusz az ige után helyezkedik el, és a neutrális mondatok is VO szórendet mutatnak (vö. Bailyn 2012), kézenfekvő az erősödő orosz nyelvi hatásnak tulajdonítani az új információ ige utáni hangsúlyosabb megjelenését, valamint a semleges VO szórendű mondatok jelenlétét.

Vizsgáltuk e folyamatok szociológiai összetevőit is, például hogyan függ össze az uráli nyelvekben az SVO mondatok aránya az orosz jelenlét és az orosz nyelvű iskolázás általánossá válásának idejével, vagy hogyan befolyásolja az SOV→SVO változást egy harmadik, SOV szórendű török nyelv jelenléte az adott területen. Kérdőíves vizsgálataink eredményei szerint a Tatárország területén élő udmurtok gyakrabban produkálják a fejevégű szórendi változatokat és rosszabbnak minősítik a fejkezdetű alternánsokat, mint az Udmurtia területén élők. Ez azt mutatja, hogy az (S)OV típusú tatár nyelv lassítja az udmurt szórendi típusváltásának folyamatát.

E kérdésekkel foglalkozik Asztalos Erika PhD disszertációja (Asztalos 2017), továbbá számos megjelent vagy megjelenés alatt álló tanulmányunk (Asztalos 2016, Asztalos, Brattico, Sahkai, Surányi 2017, Asztalos megjelenés alatt, Asztalos, Gugán, Mus 2017).

1.2. A véges alárendelés térnyerése

Hawkins (2001) elmélete szerint korreláció van az SOV szórend és a nem-véges alárendelés, valamint az SVO szórend és a kötőszóval bevezetett véges alárendelés között. Az elmélet azt jósolja, hogy az uráli nyelvekben végbemenő SOV→SVO eltolódás a véges alárendelés térnyerésével jár együtt.

E jelenséget a vonatkozó mondatokkal kapcsolatban vizsgáltuk legbehatóbban. Csepregi (2012) mutatta ki először, hogy a szurguti hanti nyelvben az ősi finnugor típusú, prenominális, participiális, vonatkozó elem nélküli vonatkozó mondatok mellett megjelentek a posztnominális vonatkozó mondatok is. Ezek lehetnek participiális és véges mondatok is, de bennük már kötelezően megjelenik a vonatkozó névmás. A projektumban ezeket az új típusú vonatkozó mondatokat elemeztük három uráli nyelvben, mind empirikus, mind elméleti szempontból.

Terepmunkáink eredményeképpen a Csepregi által a szurguti hantira tett megállapításokat új adatokkal árnyaltuk: megállapítottuk, hogy bizonyos beszélők elfogadják a posztnominális vonatkozó mondatokat vonatkozó elem nélkül is, illetve bizonyos beszélők nyelvhasználatában a kérdőszavaknak nem alakult ki vonatkozó használata. Ezek a beszélők csak a mutató névmásból grammatikalizálódott vonatkozó névmásokat tudják használni.

Az udmurt nyelvben elsőként vizsgáltuk a posztnominális vonatkozó mondatokat. Kiderült, hogy az udmurtban is kialakultak a posztnominális vonatkozó mondatok, és használatuk nagyban hasonlít a Csepregi által leírt hanti mintázathoz. A posztnominális vonatkozó mondatok az udmurtban szigorúan megkövetelik a kérdőszóból grammatikalizálódott vonatkozó névmást. A posztnominális participiális vonatkozó mondatokban csak a jövő idejű, deontikus modális értékű participium (-ono) használható. Ebben az udmurt hasonlít az angol és a román nyelvre: ezekben a posztnominális nem véges vonatkozó mondatoknak szintén deontikus modális olvasata van. Ez a mintázat viszont eltér a szurguti hantiban tapasztaltaktól: a szurgutiban a befejezett és korlátozottan a folyamatos melléknévi igenévvel alakult vonatkozó mondat is állhat a főnév után.

A nyenyec terepmunka során nyújtott adatokból az derült ki, hogy ebben a nyelvben nem állhatnak a vonatkozó mondatok a főnév után: itt csak a finnugor örökséget tükröző, prenominális, participiális, vonatkozó elem nélküli vonatkozó mondatok elfogadhatóak. Posztnominális helyzetben semmilyen formában nem fogadhatóak el a vonatkozó mondatok, és sem a kérdőszavak, sem pedig a mutató névmás nem grammatikalizálódott vonatkozó névmássá.

Elméleti szinten a következőket állapítottuk meg. A vonatkozó névmások megjelenésének feltétele a vonatkozó mondat posztnominális pozíciója. Vonatkozó névmások akkor sem jelenhetnek meg a prenominális vonatkozó mondatokban, amikor ezek már kialakultak a nyelvben és a posztnominális vonatkozó mondatokban opcionálisan vagy kötelezően megjelennek. Ezt a tényt azzal magyaráztuk, hogy a prenominális vonatkozó mondatoknak hiányzik a bal perifériája: ezek maximálisan TP méretű, csonkolt mondatok, melyekben nincsenek projektálva azok a bal periférikus projekciók, amelyekben a vonatkozó névmások szoktak helyet foglalni.

A posztnominális vonatkozó mondatokról azt állapítottuk meg, hogy ezek már rendelkeznek CP réteggel. Bennük éppen azért jelenhetnek meg a vonatkozó névmások, mert a vonatkozó névmások a CP rétegben foglalnak helyet, s ezekben a mondatokban már megvan ez a réteg. Azt is megállapítottuk, hogy az udmurt nyelvben a posztnominális vonatkozó mondatok (függetlenül a véges/participiális végződésétől) Rizzi (1997) modelljében osztott CP réteggel rendelkeznek: a vonatkozó névmások a FinP specifikálójában foglalnak helyet, és megelőzhetik őket topikok. A topikok közül elsősorban azok elfogadhatóak, melyek a CP rétegben generálhatóak (ún. frame-setting határozók), a CP rétegbe mozgatással kerülő topikok (pl. topikalizált alanyok és tárgyak) kevésbé elfogadhatóak; különösen a topikalizált tárgyak rontják a mondatok megítélését.

A vonatkozó névmás elé helyezett topikok megléte (mely az ómagyarban is megfigyelhető volt) fontos elméleti tanulságokkal jár: ez a jelenség cáfolja Kayne (1994) elemzését a vonatkozó mondatok szerkezetéről. Kayne javaslata az, hogy a vonatkozó mondatok főnévi feje a vonatkozó névmás specifikálójában foglal helyet. Ez a modell nem tudja megragadni azokat az adatokat, melyekben főnévi fej - topik - vonatkozó névmás a sorrend. Ezeket az adatokat viszont jól tudja kezelni a vonatkozó mondatok ún. "megfeleltetési" elemzése (matching), illetve az "emelő" (raising) elemzésnek az a változata, melyben a főnévi fej és a vonatkozó névmás nem alkotnak nyelvtani összetevőt.

A három nyelv összevetésével megállapítottuk, hogy a posztnominális vonatkozó mondatok annál elfogadhatóbbak, minél inkább elfogadható egy nyelvben a VO mint alapszórend az ősi finnugor OV szórend mellett (vagy helyett): a vizsgált nyelvek közül a nyenyecnek van a legszigorúbb OV szórendje és itt nem elfogadhatóak a posztnominális vonatkozó mondatok, míg az udmurtban az OV és a VO szinte egyenrangúak az alapszórend tekintetében, és a posztnominális vonatkozó mondatok itt a legelfogadhatóbbak. A szurguti hanti mindkét tekintetben a nyenyec és az udmurt között helyezkedik el.

A vizsgált nyelvek segítségével modellálni tudtuk, hogyan lesz az OV nyelvekre jellemző prenominális, participiális, vonatkozó elem nélküli mondatokból a VO nyelvekre jellemző posztnominális, véges, vonatkozó elemes vonatkozó mondatok. A változás meghatározott lépéssorozatban folyik le: először a vonatkozó mondat pozíciójának változása megy végbe, s ez előfeltétele annak, hogy a másik két paraméterben is változás mehessen végbe. Ez a modell segíti az indo-európai nyelvek történetében lezajlott változások jobb megértését is: ezek a nyelvek szintén végbementek egy OV -> VO változáson, de még a nyelvek írásos korszaka előtt. Modellünk alapján azt feltételezzük, hogy a véges vonatkozó mondatok ezekben a nyelvekben is az uráliban megfigyelt lépéssorozatokon mentek át: először posztnominális participiális vonatkozó mondatok alakultak ki, s a véges vonatkozó mondatok csak ezután jelentek meg.

A tárgyi alárendelés kapcsán elsősorban az ige komplementum pozícióját elfoglaló befejezett melléknévi igeneves (-m participium) formákat vizsgáltuk. Megállapítottuk, hogy ebben a pozícióban a befejezett melléknévi igeneves mondatok mindig viselnek (birtokos) egyeztető morfémát. Ezt azzal magyaráztuk, hogy az -m mondatok csonkolt mondatok: nincsen bal perifériájuk. Az -m maga egy T fej lexikalizációja. Tárgyi pozícióban úgy tudják ezek a mondatok megjelenni, ha nominalizálva vannak (egy kiejtetlen D fej által) - ennek a fonológiai reflexe az egyeztetés. A tárgyi -m mondatok szintaxisáról megállapítottuk, hogy adverbiumok és nem melléknévek módosítják őket, megjelenhet bennük negáció, tárgyuk tárgyesetet visel, a nominalizáció ellenére nem lehet őket többes számba tenni, alanyuk pedig Ablatív esetet visel. Ez utóbbi tényt azzal magyaráztuk, hogy az alany a nominalizáló D fej specifikálójába emelkedik, ahol levezetett birokosként funkcionál.

E témakörrel kapcsolatos kutatásainkról készült prezentációnkat (Dékány és Tanczos 2016) a 2016-os nemzetközi nyelvtörténeti kongresszuson mutattuk be. A témával foglalkozó tanulmányunk befejezés előtt áll.

1.3. A mondatrészi szerepek és a diskurzusszerepek szétválása

A konzervatív, szigorúan SOV szórendű uráli nyelvekben, például a hantiban az alany kötelezően topik szerepű, tehát a mondat élén egy alany-topik áll (egy úgynevezett SubjP projekció specifikálójában). Ha a mélyszerkezeti alany új információt közöl, passzivizációval kerül el az alany-topik pozícióba való kiemelését (még intranszitiv mondatokban is), például:

- (23) *kul'-na joxət-s-a*
ördög-LOC jön-PAST-PASS.3SG
'Megjelent egy ördög.'

A tárgy akár topik, akár fókusz szerepű lehet. A tárgyi fókusz az igei kifejezésen belül, közvetlenül az ige előtt marad; a tárgyi topik viszont az igei kifejezés elé, az alanyi topik mögé, egy ObjP projekció specifikálójába kerül, és ott nyelvtől/nyelvjárástól függően igei egyeztetést vált ki és/vagy tárgyragot kap. Határozói bővítmény (pl. részeshatározó, eszközhatározó) csak akkor lehet topik, ha határozóragját elveszítve tárggyá lép elő. Ilyenkor a mélyszerkezeti tárgy eszközhatározói ragot visel. Ez az úgynevezett szekundatív szerkezet Sipőcz (2015a,b) manysival kapcsolatos munkáiból vált ismertté. Mi a manysi és a hanti mellett a szamojéd nyelvekben is vizsgáltuk ezt az alternációt.

Az SOV szórend felbomlásával a mondatrészi és az információszerkezeti szerepek szétválnak; olyan topik és fókusz projekciók jönnek létre, melyek bármilyen mondatrészi szerepű összetevővel kitölthetők. E folyamat a hantiban még lényegében nem mutatható ki; a nyenyecben már megindult (ott az SOV mondat alanya már fókusz szerepű is lehet), míg az udmurtban már egészen előrehaladott. Az udmurt mondat szerkezetben a mondatkezdő, bármely összetevő számára elérhető topikpozíció mellett egy mondateleji kontrasztívtopik-pozíciót és egy mondatvégi információsfókusz-pozíciót is kimutattunk, melyek szintén bármilyen mondatrészi szerepű összetevővel kitölthetők.

E témakörrel kapcsolatos tanulmányaink: É. Kiss (2017a; megjelenés alatt), Bárány (2017) egyik fejezete.

2. Az uráli nyelvek kevésbé ismert mondattani sajátosságai

2.1. Differenciált tárgy–ige egyeztetés és differenciált tárgyjelölés

Sokat foglalkoztunk az uráli nyelvek nagy részére jellemző (és részben a magyarban is fennmaradt) differenciált tárgyjelöléssel és differenciált ige–tárgy egyeztetéssel. Sikerrel érveltünk amellett, hogy az egyes kelet-szibériai nyelvekben megfigyelt inverz egyeztetési tilalom az uráli nyelvekben is kimutatható; a differenciált tárgyjelölés rendszerének inherens része. E megszorítás valójában az inverz topikalitást tiltja (azaz, nem engedi, hogy az alanyi és tárgyi topikok szerkezeti hierarchiája ellentmondjon referenseiknek a topikalitási hierarchiában való rangsorának). Bizonyítottuk, hogy az uráliban megfigyelt jelenségkör az indo-európai nyelvekben nyomokban fennmaradt személy–eset megszorítás teljesebb rendszerét képviseli. Sokat vitatott, hogy e jelenség hogyan illeszthető be a grammatika rendszerébe; az érintett nyelveknek morfológiai, mondattani vagy pragmatikai sajátossága. Bizonyítékot találtunk rá, hogy az inverz topikalitás tilalma interfész jelenség: az alanyi és tárgyi topikok szerkezeti és pragmatikai hierarchiájának harmonizálását írja elő.

E témát is tárgyaló publikációink: É. Kiss (2017a), Bárány (megjelenés alatt).

2.2. Birtokos személyrag determináló funkcióban

Az uráli nyelvek többségére jellemző, de a tipológiában és az elméleti nyelvészetben jórészt ismeretlen jelenség a birtokos személyragok determináló funkcióban való használata. A jelenség kialakulásának és funkciójának jobb megértéséhez megvizsgáltuk a magyarban nem birtokosi szerepben használt birtokos személyragok funkcióját a régiségben és a mai magyarban. Bizonyítottuk, hogy az *-ik* partitivusi toldalék (*mindegyik melyik, a szebbik*) a többes szám 3. személyű birtokos személyrag egyik allomorfiájából grammatikalizálódott. Elemeztük a *-ja/je* toldalék nem birtokosi használatát is (*a java, a nagyja*, valamint: *harmadjára, hármásával*). Az udmurtban a tárgyrag is azonos formailag a 3. személyű birtokossal. Kimutattuk, hogy a tárgyrag is a 3. személyű birtokos személyrag grammatikalizálódásával keletkezett; először a tárgy határozottságát jelölte a differenciált tárgyjelölő rendszerben, és e funkciója értelmeződött át tárgyjelölő szereppé. Vizsgáltuk a birtokos személyragból grammatikalizálódó determináló toldalék és a prenominalis névelő

kialakulásának egymásra való hatását is. Releváns publikációink: É. Kiss (2017b,c), (Egedi 2017).

3. Uralisztika és nyelvelmélet

Projektumunknak fontos célja volt az uralisztikai és nyelvelméleti kutatások integrációja: az uráli nyelvek mondattanának a napjainkban uralkodó elméleti keretben való vizsgálata, és kurrens nyelvelméleti hipotéziseknek uráli anyagon való tesztelése. Ennek megfelelően eredményeinket mindig a generatív grammatikaelmélet keretében, tehát a nemzetközi nyelvészközösség számára is érvényes módszerekkel és terminológiával fogalmaztuk meg, és mindig törekedtünk a lehetséges elméleti következtetések levonására. Következtetéseink részint új általánosítások, részint korábbi hipotézisek megerősítései. Kontaktológiai általánosításunk, hogy a mondattani kontaktushatásnak háromféle megnyilvánulási formája figyelhető meg. A domináns nyelv akkor befolyásolhatja az anyanyelv grammatikáját, ha az anyanyelvnek van olyan szerkezete, mely egyaránt elemezhető az anyanyelv és a domináns nyelv szabályaival; ilyen esetben az idegen nyelvi hatás tulajdonképpen transzlingváris újraelemzés. Funkcionális elem, például kötőszó idegen nyelvből való kölcsönzése során az idegen lexikális elem magával hozhatja szerkezeti, projekciós sajátosságait (l. Longobardi (2001) inertia-elvét). Végül lehetséges egy absztrakt szintaktikai oppozíció (például a szisztematikus imperfektív–perfektív megkülönböztetés) kölcsönzése és eltérő nyelvi eszközökkel (például egy akcióminőség-jelző képző átértelmezésével) való megvalósítása.

Vizsgálataink új adatokkal támasztották alá a Final-over-Final hipotézist, és megmutatták az univerzálé magyarázatát is. A FOF (Biberauer et al. 2014, Sheehan et al. 2017) azt mondja ki, hogy egy kiterjesztett projekción belül egy fejező réteg csak fejező réteget dominálhat. Erre a történeti megfigyeléseink szolgálnak magyarázattal: a nyelvi változások a szerkezet legfelső szintjén kezdődnek, és fokozatosan terjednek az alsóbb szintekre.

Eredményeink megerősítették a változások ciklikusságára vonatkozó hipotézist is (van Gelderen 2011).

Tanulmányaink elméleti relevanciáját tanúsítja, hogy előadásainkat a legnagyobb presztízsű nemzetközi konferenciákon (pl. DIGS – Diachronic Generative Syntax) is elfogadták, és több cikkünk vezető nemzetközi folyóiratban (*The Linguistic Review*), ill. vezető nemzetközi tudományos kiadó (Oxford University Press) által kiadott tanulmánykötetben van megjelenés alatt.

Sikernek tartjuk, hogy sikerült kapcsolatot kiépíteni az uráli nyelvek mondattanával generatív keretben foglalkozó orosz, amerikai stb. nyelvészekkel. Részvételükkel két műhelykonferenciát szerveztünk; ezek anyagát a Pázmány Péter Katolikus Egyetemen egyik projektumtagunk közreműködésével szerkesztett *Finno-Ugric Languages and Linguistics* című open source folyóiratban adjuk közre.

4. Adatbázis-építés

A projekt egyik célja egy annotált adatbázis létrehozása volt udmurt, tundrai nyenyec, színjai és szurguti hanti nyelvű, írott és beszélt nyelvi szövegekből, amely lehetővé teszi az uráli–orosz kontaktushatás kutatását.

4.1. Szöveggyűjtés

Annak érdekében, hogy nyomon követhessük a kisebbségi nyelvek orosz hatásra végbemenő szintaktikai változásait, különböző korokból gyűjtött szövegek feldolgozására törekedtünk.

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 színjai hanti szövegek Wolfgang Steinitz gyűjtéséből származnak az 1930-as évekből (Steinitz, 1975), míg a szurguti hanti szövegeket Heikki Paasonen gyűjtötte

1900-01-ben a Jugán folyó környékén (Vértes, 2001). A régi udmurt szövegek két forrásból származnak: egyrészt Yrjö Wichmann gyűjtéséből (Wichmann, 1901), másrészt Munkácsi Bernát 1887-es terepmunkájából (Munkácsi, 1887). A régi tundrai nyenyec szövegek forrása Toivo Lehtisalo 1911–12-es gyűjtése (Lehtisalo, 1947).

Az új szövegek különböző műfajú forrásokból származnak: az új hanti adatok lejegyzett interjúkat tartalmaznak (Pesikova-Volkova, 2013), míg az udmurt szövegek a Mynam malpanēs¹ és a Marajko² nevű blogokból származnak. A modern tundrai nyenyec adat tartalmaz újságcikkeket a Njar'jana Ngerm című újságból, valamint új gyűjtésű folklór szövegeket Labanauskas (1995) és Puškarëva-Chomič (2001) gyűjtéseiből.

A beszélt nyelvi adatok a projektrésztvevők terepmunkái során gyűjtött anyagokból áll össze. A felvételek egy részéhez készült lejegyzés, fordítás és morfológiai elemzés is.

Minden anyag elérhető a projekt weboldalán³ keresztül. Bizonyos források szöveges anyaga mellett megtalálható a morfológiailag elemzett szövegváltozat is, ami egy fix oszlopos tab-szeparált szövegfájl (tsv), amely minden tokenszintű információt tartalmaz. Emellett elérhetőek az eredeti nyelvű szöveganyag mondat szinten párhuzamosított angol, magyar, német és orosz fordításai is. A beszélt nyelvi adatok esetében elérhetőek az audiofájlok is. Továbbá minden forrásnál megtalálható egy olyan .eaf fájl, amely minden token- és mondat szintű információt tartalmaz, a hangzó anyaghoz illesztve, amennyiben van ilyen. Ezek a fájlok az ELAN programcsomaggal⁴ megnyithatók és kereshetők.

4.2. Szövegfeldolgozás

A szöveges anyag 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 publikált 1901-ben azzal a szándékkal, hogy az uralisták által használt lejegyzési rendszereket egységesítse (Setälä, 1901). 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 feldolgozott írott nyelvi adatok egy része csak nyomtatott könyv formájában volt elérhető, ezért ezeket beszkeneltük, 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, ezért az Abbyy FineReader Professional Edition⁵ mellett döntöttünk.

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.

Átírás és konverzió

A transzkripcióval szemben meg kell különböztetnünk a transliterációt, amely egy már írott formában létező nyelvi adat átírása egy másik írás- vagy jelölési rendszerre. Az adatbázisunk tartalmaz minden szöveget legalább az eredeti lejegyzésében, amelyet a nyelv dokumentálója

¹ <http://udmurto4ka.blogspot.hu>

² <http://marjamoll.blogspot.hu>

³ <http://www.nyud.hu/oszt/elmnyelv/urali/adatbazisok.html>

⁴ <https://tla.mpi.nl/tools/tla-tools/elan/>

⁵ <http://finereader.abbyy.com>

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

A régi szurguti hanti szövegeket az Ob-Ugric Database (OUDB)⁶ fejlesztői bocsátották a rendelkezésünkre, és mivel ők csak IPA-ban tették elérhetővé az anyagukat, nekünk is csak IPA-átíratunk van. A modern szurguti hanti szövegek viszont a mai cirill betűs hanti írással íródtak, amelyet először átkonvertáltunk a Csepregi Márta által alkotott és a hanti nyelvet kutatók körében széles körben használt átírásra (Csepregi, 2011), majd ebből állítottuk 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 konvertáltuk IPA-ra.

A régi tundrai nyenyec szövegek bizonyos értelemben kivételt képeznek. Lehtisalo olyan bonyolult transzkripciósi 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 (Hajdú, 1989) transzkripciója alapján készültek el. Ez utóbbi lett IPA-ra, majd cirillre konvertálva, az utóbbi a morfológiai elemzőhöz.

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 és/vagy magyar glosszát. Ezek az információk a rendelkezésre álló morfológiai elemzők kimenetéből származnak.

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 színjai hanti és az udmurt szövegek elemzésére az MTA Nyelvtudományi Intézete és a MorphoLogic Kft. által közösen fejlesztett elemzőket használtuk (Novák, 2006; Fejes-Novák, 2010), mert ezek morféma szinten szegmentált kimenetet adnak, továbbá a magyar (és a színjai hanti esetében az angol) fordítást is előállítják.

A tundrai nyenyec szövegek elemzésére a Giellatekno⁷ elemzőjét használtuk. Sajnos a negyedik nyelvre, a szurguti hantira nem tudunk elérhető morfológiai elemzőről, de azért megpróbáltunk erre a nyelvre is valamilyen automatikus támogatást nyújtani. Egy igen egyszerű memóriaalapú megoldást alkalmaztunk. 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

⁶ <http://www.oudb.gwi.uni-muenchen.de>

⁷ <http://giellatekno.uit.no>

nyelvész szakértő kézzel hozzárendelte a szófaji kódot, az inflexiók címkéket és a lemma angol fordítását. Ezzel a glosszák több mint 60%-át tudtuk automatikusan generálni.

5. Összegzés

A projektum során lényegében sikerült a kitűzött célokat megvalósítani. Eredményeink: prezentációink, kézirataink és publikációink a projektum honlapján olvashatók. A munka során a terepmunkák megszervezése, alkalmas és készséges informánsok felderítése jelentett nehézséget. Azt tapasztaltuk, hogy az informánsok jó része csak igen magas díjazásért (napi 300 euroért vagy többért) hajlandó információkat szolgáltatni. Ez, továbbá az a tény, hogy Szibéria falvaiban az utazás, szállás gyakran csak számlát adni nem tudó magánszemélyek révén oldható meg, hasonló tárgyú projektumok esetében a jövőben rugalmasabb felhasználású pénzügyi keretet tenne szükségessé.

A pályázat tanulságait felhasználva 2017. augusztusában ERC Advanced Grant iránti pályázatot adunk be.

Hivatkozások

- Asztalos E. 2016. A fejező grammatikától a fejező felé: generációs különbségek a mai udmurt beszélőközösségben a szórendhasználat és -megítélés terén. In: É. Kiss K., Hegedűs A. és Pintér L. szerk. *Nyelvelmélet és kontaktológia* 3, 124–154. Budapest: Szent István Társulat.
- Asztalos E. 2017. Szórendi típusváltás az udmurt nyelvben. Doktori értekezés.
- Asztalos E. megjelenés alatt. Focus positions in Udmurt. *Finno-Ugric Languages and Linguistics* (<http://full.btk.ppke.hu>).
- Asztalos E., P. Brattico, H. Sahkai, B. Surányi 2017. Information structure. Megjelenik: A. Tamm és A. Vainikka szerk. *Uralic Syntax*. Cambridge: Cambridge University Press.
- Asztalos E., Gugán K. és Mus N. 2017. Uráli VX szórend: nyenyec, hanti és udmurt mondat szerkezeti változatok. In: É. Kiss K., Hegedűs A. és Pintér L. szerk. *Nyelvelmélet és diakrónia* 3, 30–62. Budapest–Piliscsaba: PPKE BTK.
- Bailyn, J. 2012. *The Syntax of Russian*. Cambridge: Cambridge University Press.
- Bárány A. megjelenés alatt. *Differential object marking in Hungarian and the morphosyntax of case and agreement*. Oxford: Oxford University Press.
- Biberauer, T., A. Holmberg és I. Roberts 2014. A syntactic universal and its consequences. *Linguistic Inquiry* 45: 169–225.
- Csepregi M. 2011. *Szurguti hanti folklór szövegek*. Budapest: ELTE Finnugor Tanszék.
- Csepregi M. 2012. Participiális jelzős szerkezetek két hanti nyelvjárásban. *Nyelvtudományi Közlemények* 108: 61–94.
- Dékány É. és Tánczos O. 2016. The development of split CP in Udmurt embedded clauses. Societas Linguistica Europaea – 49th Annual Meeting. Federico II Egyetem, Nápoly.
- Egedi B. 2017. Two paths running across the same field. The grammaticalization of referential marking in Uralic languages. Talk given at *Syntax of Uralic Languages*. 27–28 June 2017. Budapest.
- É. Kiss, K. 2017a. The Person Case Constraint and the Inverse Agreement Constraint are manifestations of the same Inverse Topicality Constraint. *The Linguistic Review* 34: 365–396.
- É. Kiss, K. 2017b. Possessive agreement turned into a derivational suffix. Megjelenés alatt. In: Bartos H., M. den Dikken, Váradi T. szerk. *Boundaries Crossed: Studies at the crossroads of morphosyntax, phonology, pragmatics, and semantics*. Dordrecht: Springer.
- É. Kiss K. 2017c. Az -ik kiemelőjel eredetéről. In: É. Kiss K., Hegedűs A. és Pintér L. szerk. *Nyelvelmélet és diakrónia* 3, 75–90. Budapest–Piliscsaba: PPKE BTK.

- É. Kiss, K. megjelenés alatt. Reconstructing Proto-Ugric and Proto-Uralic object marking. In: T Eythorsson és J Gísli Jónsson szerk. *Syntactic Features and the Limits of Syntactic Change*. Oxford: OUP.
- Fejes L., Novák A. 2010. Obi-ugor morfológiai elemzők és korpuszok. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010), 284–291. Szegedi Tudományegyetem.
- Hajdú P. 1989. *Chrestomathia Samoiedica*. Budapest: Tankönyvkiadó.
- Hawkins, J. A. 2001. Some issues in a performance theory of word order. In: A. Siewierska szerk. *Constituent Order in the Languages of Europe*, 729-782. Berlin: Mouton de Gruyter.
- Kayne, R. 1994. *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.
- Kemenade, A. van és T. Milicev 2012. Syntax and discourse in Old English and Middle English word order. In D. Jonas és S. Anderson szerk. *Grammatical Change: Origins, Nature, Outcomes*. Oxford: OUP.
- Kiparsky, P. 1996. The Shift to Head-Initial VP in Germanic. In: H. Thrainsson, J. Peter és S. Epstein szerk. *Comparative Germanic Syntax*. Dordrecht: Kluwer.
- Lehtisalo, Toivo (1947): *Juraksamojedische Volksdichtung*. Suomalais-Ugrilainen Seura, Helsinki.
- Longobardi, G. 2001. Formal syntax, diachronic Minimalism, and etymology: the history of French *chez*. *Linguistic Inquiry* 32: 275-302.
- Munkácsi B. 1887. *Votják népköltészeti hagyományok*. Budapest: Magyar Tudományos Akadémia.
- Novák A. 2006. Morphological Tools for Six Small Uralic Languages. *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, 925–930. Genoa.
- Pesikova, A. S. és A. N. Volkova 2013. *Skazki, rasskazy s reki Ljamy: Fol'klorn'j sbornik na jazyke surgutskich chanty*. Chanty-Mansijsk: Jugrafika.
- Puškarëva, E. T. és L. V. Chomič 2001. *Fol'klor nencev*. Novosibirsk: Nauka.
- Rizzi, L. 1997. The Fine Structure of the Left Periphery. In: L. Haegeman szerk. *Elements of Grammar*. Dordrecht: Kluwer.
- Setälä, E. N. 1901. Über die Transkription der finnisch-ugrischen Sprachen. *Finnisch-ugrische Forschungen* I., 15–52.
- Sheehan, M., T. Biberauer, A. Holmberg & I. Roberts 2017. *The Final-Over-Final Condition*. Cambridge, Mass.: MIT Press.
- Sipőcz K. 2015a. Ditransitivity in the Ob-Ugric languages. In: H Mantila et al. eds. *Congressus Duodecimus Internationalis Fenno-Ugristarum, Plenary Papers*, 133-159. University of Oulu.
- Sipőcz, K. 2015b. Tárgyjelölés a manyisi ditranzitív szerkezetekben. In: K É. Kiss, A Hegedüs & L Pintér eds. *Nyelvelmélet és dialektológia* 3, 234-245. Budapest: Szent István Társulat.
- Steinitz, W. 1975. *Ostjakologische Arbeiten*. Akadémiai Kiadó, Budapest.
- Taylor, A & S Pintzuk 2012. Rethinking the OV/VO alternation in Old English: the effect of complexity, grammatical weight and information status, In: T Nevalainen & E Closs Traugott eds. *The Oxford Handbook of the History of English*, 1199-1213. Oxford: OUP.
- Vértes, E. 2001. Heikki Paasonens surgutostjakische Textsammlungen am Jugan. Neu transkribiert, bearbeitet, übersetzt und mit Kommentaren versehen von Edith Vértes. In: *Mémoires de la Société Finno-Ougrienne*, 240, 67-92. Helsinki: Suomalais-Ugrilainen Seura.
- Wichmann, Y. 1901. Wotjakische sprachproben II. Sprichwörter, Rätsel, Märchen, Sagen und Erzählungen. In: *Journal de la Société Finno-Ougrienne* XIX, 1-200. Helsinki: Suomalaisen Kirjallisuuden Seuran Kirjapainon Osakeyhtiö.

