A szó elszáll, az írás megmarad? Nyelvtechnológiai eszközök a déli manysi nyelvre

Szilágyi Norbert[1], Horváth Csilla[2], Vincze Veronika[3,4], Nagy Ágoston[2]

[1]Szegedi Tudományegyetem, Finnugor Tanszék
[2]Szegedi Tudományegyetem, Angol-Amerikai Intézet
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[2]Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék
e-mail:{norbertszilagyi91,naj.agi,nagy.agoston.1}@gmail.com
vinczev@inf.u-szeged.hu

Kivonat A cikkben a veszélyeztetett nyelvek digitalizációját célzó törekvésekkel összhangban készített manysi nyelvű, elsősorban a már kihalt déli nyelvjáráshoz készített nyelvtechnológiai eszközöket és forrásokat mutatjuk be. A manysi nyelv demográfiai helyzetének, valamint nyelvjárási felosztásának felvázolása után a 19. és 20. század fordulóján dolgozott két nyelvész kutatásainak eredményeként megjelent szövegekből készített, morfológiailag annotált déli manysi korpuszt ismertetjük. Kitérünk az Univerzális Dependencia projekt által kifejlesztett morfológiai és szintaktikai címkekészlet felhasználásával készített morfológiai és függőségi annotációra (mind az északi, mind a déli manysira), illetve az elkészült digitális déli manysi – magyar szótárra is. Elegendő mennyiségű tanításra használható annotált adat birtokában gépi tanulási módszerek alkalmazását tervezzük modern manysi szövegek automatizált szófaji egyértelműsítésére és függőségi elemzésre.
Kulcsszavak: manysi, uráli, morfológia, korpusz, szintaxis

1. Bevezetés

A manysi nyelvet (Nyugat-Szibériában beszélt uráli nyelv) a kutatók lényegében a szakirodalomban megjelenő első említései óta (pl. [1]) veszélyeztettnek tekintik. A századok során állapota csak fokozottan sérülékenyebbé vált, négy nyelvváltozatából kettő (a déli és a nyugati) kihalt, a harmadik (a keleti nyelvjáráscsoport) kihalásközeli állapotban van, a negyedik (az északi nyelvjáráscsoport) pedig veszélyeztetett. A 20. század utolsó évtizedeiben végbement gyors gazdasági és társadalmi változások eredményeként a még beszélt nyelvjárások esetében drasztikusan felgyorsult a nyelvcsere folyamata, a legutóbbi két összoroszországi népszámlálás adatai alapján a manysi beszélők száma mintegy a kétharmadával csökkent.

Bár a manysi nyelv és kultúra presztízse emelkedik, a manysi nyelv multietnikus, többnyelvű környezetében csak korlátozott szerepet tölt be, nagy hatással van használatára a hagyományos életmód eltűnése és az intenzív urbanizáció. A városi életmód ugyanakkor új nyelvhasználati színterek, és a nyelvelsajátítást,

kulturális örökség megőrzését segítő új eszközök létrehozását is lehetővé teszi. A kulturális örökség még elérhető elemeinek összegyűjtése, megőrzése iránti igény az 1980-as években, a manysi nyelvi revitalizációt és a városi manysi gyerekek örökségnyelvi oktatását célul kitűző kezdeményezések az 1990-es évek végén jelentek meg a városi manysi közösségben. Az internetkapcsolat elterjedésével a digitális környezet is olcsó és könnyen hozzáférhető potenciális nyelvhasználati színtérré vált, mely egyszerre lehet eszköze egy új beszélői közösség megteremtésének és színtere a megelőző és jelenlegi nyelvi dokumentációs projektek adatainak és eredményeinek. Az online adatbázisok és számítógépes nyelvi eszközök növelik egy nyelv beszélőközösségének nagyságát és a nyelv presztízsét, vagyis azt a két jellemzőjét, mely Kornai szerint az online nyelvi vitalitás legfontosabb fokmérője [2].

A természetesnyelv-feldolgozás területén számos projekt célozza veszélyeztetett és kihalt nyelvek dokumentációját. Az Univerzális Dependencia projekt[1] keretében függőségi fákat tartalmazó szintaktikai annotáció készült például ógörög, latin, szanszkrit és óegyházi szláv nyelvre. E törekvések nyomán projektünk a rendelkezésre álló manysi írott források digitalizációjára, ezen túl számítógépes nyelvi eszközök fejlesztésére törekszik, nem kizárólag a még beszélt, hanem a már kihalt dialektusok adatainak felhasználásával is. Manysi nyelvi korpuszt készítettünk terepmunka során gyűjtött és már rendelkezésre álló forrásokból felhasznált szóbeli és írott szövegek segítségével. A cikk beszámol az automatikus szövegfeldolgozást (pl. szófaji egyértelműsítést, szegmentálást) lehetővé tévő morfológiai és szintaktikai elemzésről, elsősorban a modern manysi szövegeket szem előtt tartva. Végül bemutatjuk a digitális szótárat, mely jelen pillanatban déli manysi – magyar nyelvpárra készült el.

2. Manysi nyelvjárások

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 [3] –, 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 összegyűjtések után vette kezdetét.

A nyugati és oroszországi irodalomban (pl. [4], [5], [6]) egyaránt leggyakrabban használt felosztás négy nyelvjáráscsoportot különböztet meg. Az északi nyelvjáráscsoportba tartozó nyelvváltozatokat a Szoszva és a Ljapin folyók mentén, illetve a Berjozovói körzetben beszélik. Ez a nyelvjáráscsoport (különösen

[1] http://universaldependencies.org/

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 [7].

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 [7]. 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.

3. Manysi nyelvű korpuszok

Statisztikai gépi tanulási módszerrel működő eszközök készítése során az annotált adatbázisok kiemelkedő fontossággal bírnak a természetes nyelvek feldolgozása területén, ezért áttekintjük a manysi nyelvre rendelkezésre álló korpusztípusokat.

Tudomásunk van egy, az egyetlen manysi sajtótermék szövegei felhasználásával épített északi manysi korpuszról [8]. A korpusz 520 000 tokent tartalmaz, XML formátumban. A korpuszt bemutató cikk a digitalizációt nehezítő tényezők között sorolja fel a jelentésmegkülönböztető szerepű hosszú magánhangzók jelölése okozta helyesírási és tipográfiai problémákat. Terepmunkánk során a mindennapi élet legkülönbözőbb területeihez tartozó témákban rögzítettünk északi manysi interjúkat, az adatbázis mintegy tíz órányi spontán beszédet tartalmaz. A felvételek lejegyzése jelenleg is tart, ezután a szövegek az északi manysi nyelvű korpuszt fogják bővíteni.

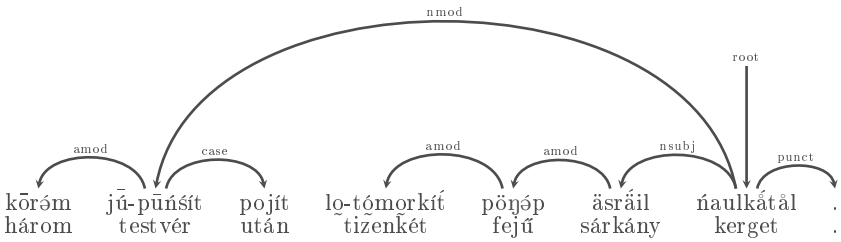
Kihalt dialektus lévén a déli manysi korpuszt nem kortárs szövegekből vagy interjúkból építettük, hanem két nyelvész, Munkácsi Bernát 1888-ban és 1889-ben, valamint Artturi Kannisto 1903-ban és 1904-ben, vagyis több mint száz évvel ezelőtt gyűjtött szövegmutatványait használtuk fel. Mindketten hosszú, átfogó, a nyelvi adatok mellett a mindennapi életre, folklórra, anyagi kultúrára, hiedelemvilágra is kiterjedő kutatómunkát folytattak, gyűjtésük megjelent anyagának a Vogul Népköltési Gyűjtemény [9] negyedik, illetve a Wogulische Volksdichtung [10] 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[2]. Az adatbázis 2400 mondatból és 11 500 szóalakból áll, ebből 5000 külön lexikális egység. A korpusznak SIL FieldWorks Language Explorerrel (FLEx[3]) kézzel annotált, morfológiailag elemzett verziója is elkészült, az annotáció a magyar fordítással együtt elérhető a honlapon, vagyis az adatok párhuzamos korpuszként is használhatók.

[2] http://norbertszilagyi91.wixsite.com/tawdamansi
[3] http://fieldworks.sil.org/flex/

Mindegyik korpuszunk felhasználhatóságát bővítendő, az adatokat morfológiai és szintaktikai annotációval látjuk el. E célra az Univerzális Dependencia projekt által kifejlesztett morfológiai és szintaktikai címkekészletet [11] alkalmazzuk, különös figyelmet fordítva a rokon nyelvekre (pl. finnre és magyarra) kidolgozott nyelvspecifikus jegyekre. Az annotált adatok az Univerzális Dependencia projekt oldalán lesznek elérhetők. Alább az 1. és a 2. ábrán mutatunk példát egy annotált mondatra.

kō̆rə́m	jṵ̄-pṵ̄ńśít	pojít	
NUM	NOUN	ADP	
NumType=Card	Case=Nom		
	Number=Plural		
lo̯-to̯mo̯rkíť	pöŋə́p	äsrə̆̄il	ńaulkə̆̄tål
NUM	ADJ	NOUN	VERB
NumType=Card	Case=Nom	Case=Nom	Definite=Indef|Mood=Ind
	Degree=Pos	Number=Sing	Number=Sing|Person=3|Tense=Pres

1. ábra: Egy déli manysi mondat morfológiai elemzése („A tizenkét fejű sárkány kergeti a három testvért.”).



2. ábra: Egy déli manysi mondat függőségi elemzése („A tizenkét fejű sárkány kergeti a három testvért.”).

4. Lexikális erőforrások

A 19. és 20. század fordulóján tevékenykedő kutatók manysi szóanyagából összeállított szótárak csak jelentős késéssel jelentek meg [9,10], a szótárak minden, így az azóta kihalt dialektusok szóanyagát is tartalmazzák. Ezen kívül rendelkezésre állnak csak az északi manysi nyelvváltozaton alapuló, némileg modernebb szótárak [12,13] is. 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 3. pontban bemutatott korpuszban szereplő minden szót és toldalékot tartalmaz, annak magyar fordításával és morfológiai információkkal együtt. A lista az egyes elemek alapváltozatait, allomorfjait is tartalmazza. Az alábbiakban bemutatunk egy szótári alakot és egy toldalékot az adatbázisból.

äsrə̑il (n)
sárkány, ördög

-ɔ́lt (v>v)
CAUS (műveltető)
əlt

Elsőként a szótári alak szerepel, szófaji információkkal együtt ((n – főnév és v>v mint igéből igét képző toldalék), majd a magyar fordítása és a helyesírási változatokat korpuszadatok alapján. Összesen 1333 elemet tartalmaz a szótár, melyek szófaji eloszlása az 1. táblázatban látható.

Szófaj	**Elemszám**
Főnév	481
Ige	338
Határozószó	97
Melléknév	75
Névutó	55
Partikula	38
Toldalék	249

1. táblázat. Szófajok a déli manysi szótárban.

5. Nyelvtechnológiai eszközök

A manysi egy gazdag morfológiájú nyelv, tehát a hozzá készült NLP-eszközök jó morfológiai elemzőt igényelnek. Annak ellenére, hogy ez egy veszélyeztetett nyelv, néhány morfológiai elemző készült már hozzá. Az északi manysihoz ugyan van morfológiai elemző [14,15], de latin karaktereket használ és csak a Kálmán-féle Chrestomathia Vogulica [5] és Wogulische Texte [16] szövegeire, illetve Munkácsi szövegeire van optimalizálva. Ezen felül sajnos nem is nyílt forráskódú.

Az északi manysihoz Vincze és mtsai [17] már említenek egy morfológiai elemzőt, amely nem olyan rég óta érhető el a Giellatekno webhelyről[4]. Ez az elemző már a mai manysi szövegekre van optimalizálva, amelyek a cirill átírást

[4] http://giellatekno.uit.no/cgi/d-mns.eng.html

használják. Ez egy nyílt forráskódú eszköz, amelynek nyelvtani alapjait később ki szeretnénk terjeszteni a déli manysi latin betűs szövegeinek feldolgozásához. Ezen felül, mivel az északi és a déli dialektusok között vannak különbségek (pl. a déli dialektusban a zéró-morfémás jövő idő vs. jelölt jelen idő [18]), a morfológiai paradigmák átírása is szükséges, amelyre csapatunk vállalkozik. Ebben a munkafolyamatban felhasználjuk a korpusz kézzel annotált mondatait is, amelyben a szóalakok szótövekkel és affixumokkal szerepelnek, és ezt beépíthetjük az elemzőbe is.

Mint ahogy korábban említettük, jelenleg a korpusz morfológiai és szintaktikai annotációján dolgozunk. Amint lesz elegendő mennyiségű annotált adatunk, amelyet tanításra használhatunk, tervezzük gépi tanuló módszerek alkalmazását automatizált POS-taggelésre és függőségi elemzésre (parsing), elsődlegesen a modern manysi szövegek automatikus elemzését szem előtt tartva. Az így elkészülő nyílt forráskódú POS-taggerhez és elemzőhöz az elemző iránt érdeklődő személyek ingyenesen hozzáférhetnek.

6. Összefoglalás

A cikkben a veszélyeztetett nyelvek digitalizációját célzó törekvésekkel összhangban készített manysi nyelvű, elsősorban a már kihalt déli nyelvjáráshoz készített nyelvtechnológiai eszközöket és forrásokat mutattuk be. A manysi nyelv demográfiai helyzetének, valamint nyelvjárási felosztásának felvázolása után a 19. és 20. század fordulóján dolgozott két nyelvész kutatásainak eredményeként megjelent szövegekből készített, morfológiailag annotált déli manysi korpuszt ismertettünk, továbbá az északi manysira elkészült korpuszoktól is szót ejtettünk. A korpuszba emelt déli manysi szövegek eredetileg a kutatók gyűjteményes köteteiben jelentek meg évtizedekkel ezelőtt, így már a digitalizálásukkal is elérhetőek váltak úgy nyelvészettel, természetes nyelvek feldolgozásával foglalkozó tudományos körök, mint az őshonos népek kulturális öröksége iránt érdeklődő szélesebb közönség számára is. Ezen felül a szövegben található néprajzi adatok a manysik, vagy általában az oroszországi őshonos kisebbségi népek iránt érdeklődő etnográfusok, folkloristák érdeklődésére is számot tarthatnak.

Kitértünk az Univerzális Dependencia projekt által kifejlesztett morfológiai és szintaktikai címkekészlet felhasználásával készített morfológiai és szintaktikai annotációra, illetve az elkészült digitális déli manysi – magyar szótárra is. Elegendő mennyiségű tanításra használható annotált adat birtokában gépi tanulási módszerek alkalmazását tervezzük (elsősorban modern, északi nyelvjárásban íródott) manysi szövegek automatizált szófaji egyértelműsítésére és függőségi elemzésre.

Reméljük, hogy mind a korpuszunk, mind a készített eszközök segítséget nyújtanak a nyelvük története iránt érdeklődő manysiknak és a veszélyeztetett, kihalásközeli és kihalt nyelvek iránt érdeklődő NLP-kutatók számára egyaránt.

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. Munkácsi, B.: Nyelvészeti tanúlmányútam a vogulok földjén [My Linguistic Fieldwork in the Land of Voguls]. Budapesti szemle **17**(60) (1889) 383–408
2. Kornai, A.: Digital language death. PLoS ONE **8**(10) (2013) e77056
3. Sipőcz, K.: www.perepis2002.ru. Finnugor Világ **10**(2) (2005) 23–27
4. Riese, T.: Vogul. Number 158 in Languages of the World/Materials. Lincom Europa, München - New Castle (2001)
5. Kálmán, B.: Chrestomathia Vogulica. Tankönyvkiadó, Budapest (1976)
6. Rombandeeva, E.I., Vakhruševa, M.P.: Mansijskij jazyk. Prosveščenije, Leningrad (1984)
7. Keresztes, L.: Mansi. In Abondolo, D., ed.: The Uralic languages, Routledge (1998) 387–427
8. Horváth, C., Szilágyi, N., Nagy, A., Vincze, V.: Language technology resources and tools for Mansi: an overview. In: Proceedings of the Third International Workshop on Computational Linguistics for Uralic Languages, St. Petersburg, Russia (2017)
9. Munkácsi, B., Kálmán, B.: Wogulisches Wörterbuch. Akadémiai Kiadó, Budapest (1986)
10. Kannisto, A.: Wogulisches Wörterbuch. Kotimaisten Kielten Keskuksen Julkaisuja, Helsinki (2013)
11. Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D.: Universal Dependencies v1: A Multilingual Treebank Collection. In: Proceedings of LREC 2016. (2016)
12. Rombandeeva, E.I.: Russko-mansijskij slovar'. Mirall, Sankt-Peterburg (2005)
13. Rombandeeva, E.I., Kuzakova, E.A.: Slovar' mansijsko-russkij i russko-mansijskij. Prosvešenije, Leningrad (1982)
14. 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
15. Fejes, L., Novák, A.: Obi-ugor morfológiai elemzők és korpuszok. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010), Szegedi Tudományegyetem (2010) 284–291
16. Kálmán, B.: Wogulische Texte mit einem Glossar. Akadémiai Kiadó, Budapest (1976)
17. Vincze, V., Nagy, A., Horváth, C., Szilágyi, N., Kozmács, I., Bogár, E., Fenyvesi, A.: FinUgRevita: Developing Language Technology Tools for Udmurt and Mansi. In: Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages, Tromsø, Norway (2015)
18. Honti, L.: System der paradigmatischen Suffixmorphemen des wogulischen Dialektes an der Tawda. Akadémiai Kiadó, Budapest (1975)