Záróbeszámoló

Finnugor nyelvű közösségek nyelvtechnológiai támogatása online tartalmak létrehozásában

(FNN 107885)

2018. március 28.

1. Bevezetés

Az elmúlt évtizedekben zajló digitális forradalom egyik hatása, hogy az információs igény nagy része online elérhető közösségi forrásokból származik, olyanokból, mint a Wikipédia[1] és a Wiktionary[2], melyeket önkéntesek ezrei szerkesztenek. A kommunikációs technológia a mindennapi életünkben, sőt a személyes szféránkban is egyre fontosabb szerepet játszik. A nyelv szerepe ezekben az új helyzetekben és alkalmazásokban kulcsfontosságú, hiszen a nyelv a közvetítő közeg. Ezért is kiemelten fontos, hogy a nyelv gyorsan adaptálódjon az új helyzetekhez, különben a kihalás veszélye fenyegeti. Az általánosságban elmondható, hogy a kommunikációs technológia vívmányaira a nyelvi közösségek nagyon érzékenyek, de a regionális és kisebbségi nyelveket beszélők közössége a lehető legérzékenyebb, vagyis az ő esetükben fordulhat elő a legkönnyebben, hogy a veszélyeztetett nyelvek csoportjába kerülnek.

Hagyományosan veszélyeztetett nyelveknek azokat a nyelveket szokták nevezni, amelyeknek kevés a beszélője, ők is az idősebb generációba tartoznak, így a beszélők száma folyamatosan csökken, és a nyelvhasználat területe az informális keretek felé tolódik. Kornai [2013] és Ács et al. [2017] a fenti tényezők mellett – többek között – a nyelvtechnológiai (és tágabban az infokommunikációs) eszközök használatát és a webes tartalmak előállításának ütemét is beleveszi a nyelvek állapotának kiértékelésébe.

A nyelvtechnológia ebben a kontextusban egyfajta támogató technológiaként tud működni: a szabad és nyilvános nyelvhasználatot támogatva, a nyelvi határokat ledöntve segíti a kommunikációt [Simon et al., 2012]. Nyelvtechnológiai alkalmazások és erőforrások viszont leginkább a széles körben használt nyelvekre léteznek. Ennek legfőbb oka az, hogy ezeken a nyelveken érhető el digitális szöveges tartalom. A kisebb, veszélyeztetett nyelvek ebből a szempontból is hátrányban vannak, hiszen hozzáférhető digitális tartalom híján nyelvtechnológiai eszközöket is sokkal nehezebb rájuk fejleszteni.

Ebből a helyzetből kiindulva a projekt távlati célja az volt, hogy kisebbségi finnugor nyelvű közösségeket segítsen a digitális revitalizációban azáltal, hogy online tartalmakat hoz létre az adott nyelveken. A projekt során több veszélyeztetett finnugor és néhány széles körben használt viruló

[1] `https://www.wikipedia.org/`

[2] `https://www.wiktionary.org/`

nyelvre állítottunk elő ún. protoszótárakat, amelyeket különféle nyelvtani információkkal gazdagítva feltöltöttünk a Wiktionarybe.

Hat kisebbségi finnugor nyelvvel dolgoztunk forrásnyelvként, melyek a következők: komi-permják, komi-zürjén, udmurt, mezei mari, hegyi mari és északi számi. A fordítások célnyelvének négy olyan, több beszélővel rendelkező nyelvet választottunk, melyek ezen nyelvközösségek szempontjából fontos szerepet játszanak. Ezek az angol, a finn, a magyar és az orosz. Egy forrásnyelv négy célnyelvvel tud párt alkotni, így összesen 24 nyelvpár számára állítottunk elő kétnyelvű szótárakat.

2. A kitűzött feladatok

A projekt globális célkitűzése az volt, hogy nyelvtechnológiai támogatást nyújtson kisebbségi finnugor digitális nyelvi közösségeknek azáltal, hogy online tartalmakat állít elő – így támogatva ezeket a közösségeket a digitális revitalizációban és hozzájárulva a nyelvi sokszínűség fennmaradásához. Ezt olyanformán terveztük megvalósítani, hogy protoszótárakat állítunk elő a fent nevezett nyelvpárokra, majd ezeket feltöltjük a Wiktionarybe.

A protoszótárak előállításához a párhuzamos és összevethető korpuszokon alapuló sztenderd automatikus szótárépítési módszereket terveztük használni. A korpuszok anyagát egyrészt a University of Helsinki Language Corpus Server (UHLCS)[3] többnyelvű szövegeiből, másrészt a Wikipédia adott nyelvű szócikkeiből terveztük összeállítani. A kutatási tervben hangsúlyoztuk, hogy az összevethető korpuszok szótárépítésre való használata önmagában egy kutatás alatt álló terület, így számos kísérletet terveztünk ebben a témában lefolytatni.

A projekt kezdetekor már létező szövegfeldolgozó eszközöket terveztük használni az adott nyelvű szövegek mondatokra és tokenekre bontásához, valamint morfológiai elemzéséhez. Az ilyenformán előfeldolgozott szöveg volt szánva az automatikus szótárépítési metódus bemenetének, amelyhez szükséges minden nyelvpárra egy ún. magszótár is. Ez utóbbiakat az UraloNet [Bakró-Nagy et al., 2015] etimológiai adatbázisból terveztük kinyerni.

Az automatikus szótárépítés sztenderd megközelítése párhuzamos vagy összevethető korpuszokból történő kontextushasonlóság-számításon alapul [Fung and Yee, 1998; Rapp, 1995]. Az volt a tervünk, hogy a Héja and Takács [2012] által leírt metódust követve először az összevethető korpuszokból nyerjük ki a párhuzamos szövegrészeket, majd automatikus mondat- és szóillesztő eszközökkel állítjuk elő a lehetséges fordítási párokat.

Az ilyen módon előállított szótárakat terveztük feltölteni a Wiktionarybe – nem csak a Wiktionary saját markup nyelvét követve, hanem olyan formátumban is, amely lehetővé teszi, hogy a szemantikus web részévé váljanak az újonnan előállított lexikai erőforrások.

3. Az elért eredmények

3.1. Korpuszépítés

A tervek szerint haladva a munkát korpuszépítéssel kezdtük. Míg a párhuzamos és összevethető korpuszok a szótárépítéshez szükségesek, az egynyelvű korpuszok tanítóanyagként tudnak szolgálni a tokenizáló és mondatra bontó alkalmazások számára. Valódi párhuzamos szövegnek csak a biblia- és regényfordítások, a szoftverdokumentációk és az olyan hivatalos dokumentumok, mint például az

[3] `http://www.ling.helsinki.fi/uhlcs/`

Egyetemes Emberi Jogok Nyilatkozata tekinthetők. Ezeken felül forrásként használtuk még Finnország és Norvégia egyes hivatalosan kétnyelvű régióinak weboldalait is. Az összevethető korpuszok elsődleges forrása a Wikipédia volt. További módszer összevethető korpuszok építésére az azonos téma köré szerveződő alkorpuszok felhasználása, vagyis olyan egynyelvű szövegek letöltése különböző nyelveken, amelyek azonos tárgykörhöz tartoznak [Fung and Yee, 1998]. Az egynyelvű korpuszokat felépítő anyagokat különféle weboldalakról töltöttük le, így ezek témában igen változatosak (pl. irodalmi szövegek, hírek, személyes blogok, hivatalos szövegek). A szöveggyűjtésről és az összegyűjtött korpuszok méretéről részletesen beszámoltunk korábbi cikkeinkben: Benyeda et al. [2015]; Simon et al. [2015].

A szótárelőállítás további lépéseihez elengedhetetlenül szükséges az összegyűjtött szövegek minél pontosabb alapszintű nyelvi feldolgozása, vagyis a tokenizálás, a mondatra bontás, a morfológiai elemzés és egyértelműsítés, mivel az ezen feldolgozási szakaszokban bekövetkezett hibák jelentős problémákat okozhatnak a magasabb feldolgozási szinteken, illetve a szótárépítésben. Az a tény azonban, hogy az általunk vizsgált kisebbségi finnugor nyelvek kevés erőforrással rendelkeznek, igen nagy mértékben korlátozta a lehetőségeinket. A projekt kezdetekor csak a Giellatekno[4] keretrendszerében létezett néhány, fejlesztés alatt álló szövegfeldolgozó eszköz ezekre a nyelvekre, így olyan módszerekkel és szoftvercsomagokkal kísérleteztünk, amelyek eredetileg nagy nyelvekre lettek kifejlesztve. Például a két mari vagy a két komi nyelvet keverten tartalmazó szövegek esetében a nyelvek megkülönböztetésére a Blacklist Classifiert[5] használtuk, amely 97,47%-os pontossággal szűrte a komi-zürjén és komi-permják, 96,77%-os pontossággal pedig a mezei és hegyi mari nyelveket. A mondatra bontáshoz és tokenizáláshoz az Apache OpenNLP[6] mondatra bontó és tokenizáló moduljait használtuk, melyek 98% feletti F-mértékkel teljesítettek. Ez a teljesítmény nagyrészt a mondatra bontó által használt rövidítésszótárnak köszönhető, melyet a Wiktionary orosz rövidítéslistája alapján állítottunk össze.

A szótárépítés szempontjából kiemelten fontos a lemmatizálás kérdése, hiszen a szótárak a szavak ún. szótári alakját tartalmazzák. Ha egy szöveg minden egyes tokenjének a tövét szeretnénk megkapni, akkor először is szükség van egy morfológiai elemzőre. Ennek a kimenete azonban még nem elég, mert az elemző kiadja az összes lehetséges elemzést. Ezek közül kontextuális jegyek alapján lehet kiválasztani a megfelelőt, amihez szükség van egy nyelvmodellre, aminek a betanítását egy morfológiailag annotált szövegen lehet megtenni. Morfológiai elemző létezik északi számi[7], udmurt és komi-zürjén[8], valamint hegyi mari[9] nyelvekre, a projektnek ebben a fázisában azonban nem volt még elérhető korpusz, amin tanítani lehetett volna egy felügyelt gépi tanuló rendszert. A projekt kezdete óta némileg javult a helyzet: tudomásunk szerint jelenleg a Saami International Korpus (SIKOR)[10] 28,41 millió tokennyi északi számi szöveget tartalmaz morfológiai és szintaktikai annotációval ellátva. A szövegfeldolgozás lépéseiről és nehézségeiről korábbi cikkeinkben számolunk be: Benyeda et al. [2015]; Simon et al. [2015].

[4] http://giellatekno.uit.no/
[5] https://bitbucket.org/tiedemann/blacklist-classifier/wiki/Home
[6] https://opennlp.apache.org/
[7] http://giellatekno.uit.no/cgi/index.sme.eng.html
[8] http://www.morphologic.hu/urali/
[9] http://www.univie.ac.at/maridict/site-2014/morph.php
[10] http://gtweb.uit.no/korp

3.2. Szótárépítés

A kétnyelvű szótárak kézzel való előállítása időigényes feladat, mely nagy fokú hozzáértést és precizitást igényel. Ezért a kevés beszélővel rendelkező nyelvek számára nem gazdaságos ez a fajta szótárépítés. Komplett kétnyelvű szótárak teljesen automatikusan történő előállítását a jelenlegi technológia nem teszi lehetővé, ezért automatikus módszerekkel ún. protoszótárakat hoztunk létre, melyek fordítási jelölteket tartalmaznak, és kézi ellenőrzést igényelnek.

Az automatikus szótárépítés szakirodalmi feltárása után számos módszerrel kísérleteztünk. Az összevethető korpuszokból történő szótárépítési metodológia sztenderd megközelítése kontextusvektorok hasonlóságát méri a két vizsgált nyelvre [Fung and Yee, 1998; Rapp, 1995], aminek a lépései a következők: kontextusvektorok létrehozása és fordítása, a forrás- és a célnyelvi vektorok összehasonlítása és a fordítási jelöltek rangsorolása valamilyen hasonlósági metrika alapján. Ehhez a módszerhez szükség van egy ún. magszótárra, amelynek használatával újabb fordítási párokat nyerhetünk ki a szövegekből. A módszer hátránya, hogy a teljesítménye erősen függ a magszótár, a kontextus és a korpusz méretétől, valamint a választott hasonlósági metrikától is. Ezeken felül számos újabb módszert is alkalmaztak nem párhuzamos korpuszokból történő fordítási párok kinyerésére [Hazem and Morin, 2012; Tamura et al., 2012; Vulić and Moens, 2012]. Az elmúlt években a forrás- és célnyelvi szavakat reprezentáló vektorokat jellemzően szóbeágyazást alkalmazó módszerekkel nyerik ki [Vulić and Moens, 2015]. Mivel az általunk vizsgált finnugor nyelvekre nem áll rendelkezésre megfelelő méretű korpusz és szótár, alternatív módszerekkel kellett kísérleteznünk.

Szabadon elérhető, nyelvfüggetlen eszközökkel végeztünk néhány kísérletet: a `Hundict`[11] és a `Hunalign` [Varga et al., 2007] felhasználásával sikerült olyan kétnyelvű szótárakat létrehoznunk, amelyek lehetséges fordítási megfelelőket tartalmaznak a hozzájuk tartozó konfidenciaértékekkel. Az eredmények kecsegtetők lettek volna, ha a bemenet lemmatizált szöveget tartalmazott volna. De mivel a 3.1. fejezetben ismertetett helyzet miatt ez sajnos nem volt számunkra elérhető, még újabb módszerek felé fordultunk.

A protoszótárak előállításához két, közösség által épített nyelvi erőforrást használtunk fel, a Wikipédiát és a Wiktionaryt. A Wikipédia többféle módon is felhasználható kétnyelvű szótárak létrehozására. Mi Erdmann et al. [2009] és Mohammadi and GhasemAghaee [2010] módszerét követve kétnyelvű szótárakat hoztunk létre Wikipédia-címszópárokból a nyelvközi linkek segítségével. A Wikipédia mellett a Wiktionary egy másik, szintén nyílt, közösség által szerkesztett tudásbázis, amely forrásul szolgálhat kétnyelvű szótárak létrehozásához. Bár a Wiktionary elsősorban emberi felhasználásra készült, a benne található adatok kinyerése bizonyos fokig automatizálható. Ács et al. [2013] minden címszóhoz tartozó fordítási megfelelőt kinyert a szócikkekben található fordítási táblákból. Az általuk fejlesztett `Wikt2dict`[12] eszközzel feldolgoztuk az angol, finn, orosz és magyar Wiktionary-oldalakat, így szinte minden szóban forgó nyelvpárra sikerült fordítási párokat kinyernünk. Ács [2014] a szópárok halmazát újabbakkal bővítette úgy, hogy háromszögeléssel új kapcsolatokat hozott létre a már meglévő fordítási párokból. A háromszögelés azon a feltételezésen alapul, hogy két elem nagy valószínűséggel fordításpár abban az esetben, ha mindkettő egy harmadik nyelv szavának fordítása. A `Wikt2dict` háromszögelési technikájával protoszótárainkat tovább tudtuk bővíteni. A szótárépítés különféle módszereiről és kihívásairól bővebb leírást tartalmaznak cikkeink: Benyeda et al. [2015]; Simon et al. [2015]; Benyeda et al. [2016].

[11] `https://github.com/zseder/hundict`
[12] `https://github.com/juditacs/wikt2dict`

3.3. Kiértékelés

A kiértékelésbe a fent leírt alternatív módszerek mellett olyan szótárakat is bevontunk, amelyeket az Opus korpuszból [Tiedemann, 2009] töltöttünk le. A letöltés idejében az általunk vizsgált nyelvpárok közül csak az északi számi–{angol, finn, magyar} nyelvpárokra találtunk szótárakat. Ezek a szótárak a 3.2. fejezetben említett sztenderd szótárépítő módszerekkel készültek, amelyekről azt feltételeztük, hogy nem lesznek megfelelőek az általunk vizsgált kevés erőforrással rendelkező nyelvek esetében. A kapott eredmények ezt alátámasztják, hiszen 27,57%-os pontosságával ez a módszer adta a legrosszabb eredményt.

A kiértékelés első lépéseként a különböző módszerekkel előállított protoszótárakat célnyelvenként összevontuk, majd az ismétlődő szópárokat kiszűrtük. Az összevont szótárak kézi kiértékelését az adott nyelvek anyanyelvi beszélői és nyelvész szakértői végezték. A validátorok nagy része Oroszországban él, ezért a velük való együttműködés céljából többször mentünk terepmunkára, illetve láttuk őket vendégül. A validátorok számára részletes instrukciókkal ellátott útmutatót dolgoztunk ki magyar, angol és orosz nyelven, valamint bevezettünk olyan kategóriákat, melyek azt jelzik, hogy egy adott szópár megfelel-e ezeknek az instrukcióknak, vagy sem. Ezek a kategóriák képezték a kiértékelés alapját.

Az automatikusan létrehozott protoszótárak kézi kiértékelése és javítása több célt is szolgált. Egyrészt lehetőséget adott az általunk használt szótárépítési módszerek összehasonlítására, másrészt megadta azoknak a szópároknak a számát, amelyeket feltölthettünk a Wiktionarybe. Ami az előbbit illeti: az általunk kipróbált és kiértékelt szótárépítési módszerek közül a Wiktionary-alapú módszerek bizonyultak a legpontosabbnak, de a Wikipédia-címszópárokból építkező módszer is jól teljesített. Várakozásainknak megfelelően a sztenderd szótárépítési metódussal előállított protoszótárak voltak a legkevésbé hasznosak. Ami pedig a Wiktionarybe kerülő szócikkek számát illeti: a célnyelvi Wiktionarykben szereplő forrásnyelvi lexikai elemek számát megsokszoroztuk. A kiértékelés eredményeit részletesen ismertetjük a következő cikkeinkben: Simon and Mittelholcz [2017]; Ferenczi et al. [2018a,b]; Simon et al. [2018a,b].

3.4. Wiktionary-szócikkek generálása

A Wiktionary-szócikkek alapjául a validált protoszótárak szolgálnak. Például az északi számi–finn nyelvpár esetén az északi számi szó a finn Wikisanakirja egy új címszava, míg a finn megfelelője a definíció lett. A szócikkek kötelező és kiegészítő elemei is teljesen automatikusan készültek.

Minden Wiktionary-kiadásnak megvannak a maga szabályai a szócikk felépítését illetően. Nem csak a formára vonatkozó szabályokat írják le, hanem azt is, hogy milyen nyelvtani információkat kell tartalmaznia egy szócikknek. A négy célnyelvi Wiktionary leírásai alapján sikerült egy olyan általános felépítést meghatározni, mely tartalmazza a címszót, a címszó nyelvét, annak szófaját és a fordítási megfelelőjét. Ezen kötelező elemek közül csak a szófaji címkék hiányoznak a szótárainkból, ezért ezeket morfológiai elemzők és különféle egyértelműsítő heurisztikák segítségével állítottuk elő. Kiegészítő információként IPA-átírást is rendeltünk a forrásnyelvi szavakhoz.

A teljesen automatikusan generált szócikkek feltöltése is automatikusan zajlott. Erre a célra a MediaWiki `Pywikibot`[13] nevű keretrendszerét használtuk, ami sima szöveges állományokból generál wiki oldalakat, amiket automatikusan feltölt a megadott nyelvű Wiktionarybe. A botok használatát mindazonáltal erősen szabályozza a Wiktionaryk szerkesztősége, ezért a projekt végére csak a magyar és a finn Wiktionary-kiadásokba tudtuk feltölteni az újonnan előállított cikkeinket. Az angol

[13] `https://www.mediawiki.org/wiki/Manual:Pywikibot`

és az orosz kiadások szerkesztőivel tárgyalunk a botengedély ügyében. A Wiktionary-szócikkek automatikus előállításáról, a szófaji címkék hozzárendeléséről és mindezek kiértékeléséről bővebb leírás található a Simon et al. [2018b] és Ferenczi et al. [2018b,a] cikkeinkben.

A szótári elemek a Wiktionary különböző nyelvű változataiban összekapcsolhatók, az interwiki linkek pedig a Wikipédia felé biztosítják az átjárást. Ez lehetővé teszi, hogy a nyelvközösségek gazdag lexikai anyaghoz férjenek hozzá. Szabadon elérhető online többnyelvű lexikai erőforrás a sok beszélővel rendelkező nyelvekre is kevés van – kivétel ez alól a BabelNet [Navigli and Ponzetto, 2012] és a szabadon elérhetővé tett többnyelvű wordnetek, mint például a MultiWordNet [Pianta et al., 2002] –, vagyis a megbízható lexikai erőforrások előállítása minden nyelvre kiemelten hasznos. 2018 márciusában jött ki a BabelNet legújabb, 4.0 verziója, amelynek egyik újdonsága, hogy azokból a lexikai erőforrásokból, amelyekből építkezik (Wikipédia, Wiktionary, Wikidata, GeoNames stb.), mindennap automatikusan letölti és integrálja a legfrissebb verziót. Azzal, hogy több ezer újonnan generált szócikket töltöttünk fel a Wiktionarybe, a BabelNetbe is átkerültek ezek a szavak, amin keresztül pedig a szemantikus web részévé váltak, köszönhetően annak, hogy a BabelNet RDF-fé alakít minden összegyűjtött információt.

4. Változások és eltérések

A 2. fejezetben ismertetett munkatervtől abban a tekintetben eltértünk, hogy nem a sztenderd módszereket használtuk a szótárak létrehozására, hanem alternatív módszerekkel kísérleteztünk (lásd a 3.2. fejezetet). Ezt az indokolta, hogy a sztenderd módszerek nagy mennyiségű előfeldolgozott szöveget és ún. magszótárakat igényelnek, míg az általunk vizsgált kisebbségi finnugor nyelvekre nem állt rendelkezésre se nagy méretű korpusz, se az alapszintű szövegfeldolgozó eszközök, se magszótár. Ez utóbbi létrehozására tettünk kísérleteket [Benyeda et al., 2016] – ezek kimenete végül a végső validálandó szótárakba lett beolvasztva. A szótárépítő módszerek alapos kiértékelése alátámasztotta azt a kezdeti hipotézisünket, hogy a sok erőforrással rendelkező nyelvekre kifejlesztett sztenderd szótárépítési módszerek nem jól alkalmazhatók a kevés erőforrással rendelkező kisebbségi nyelvekre, mivel a többek között [Héja and Takács, 2012] által leírt és előzetesen tervezett metódus a maga 27,57%-os pontosságával a legrosszabb teljesítményt nyújtotta.

Terveink szerint a korpuszépítéshez az UHLCS szövegeit, valamint a Wikipédiát használtuk volna (lásd a 2. fejezetet). A tervektől való eltérés ott történt, hogy az UHLCS szövegeiről kiderült, hogy az általunk vizsgált kisebbségi finnugor nyelveken elérhető szövegek nem az eredeti cirill betűs formában, hanem egy önkényes latin betűs átiratban szerepelnek, ezért nem tudtuk őket használni. Találtunk viszont más forrásokat az egynyelvű, párhuzamos és összevethető korpuszok építéséhez, ahogy azt a 3.1. fejezetben ismertettük. Az eredetileg magszótárként alkalmazni tervezett UraloNettel hasonló tapasztalatunk volt: sajnos nem tudtuk használni se erre, se más célra (például a Wiktionary-szócikkek opcionálisan alkalmazható etimológiai részének automatikus előállításához), mivel az abban felsorolt leánynyelvi alakok mind latin betűs átiratokban szerepelnek.

Az eredeti tervben társkutatóként megnevezett Héja Enikő és Kuti Judit szülési szabadságra mentek, Lendvai Piroska pedig külföldön vállalt munkát. Helyükre fiatal kutatók érkeztek: Benyeda Ivett, Ferenczi Zsanett, Koczka Péter, Ludányi Zsófia, Mittelholcz Iván, Tóth Bianka. A projekt időtartama alatt a résztvevők körében született egy BSc (programtervező informatikus), egy MA (digitális bölcsész) és két PhD (magyar nyelvészet és elméleti nyelvészet) fokozat. A projekt témájából kinőtt egy MA szakdolgozat is, ami ebben a félévben kerül leadásra.

5. Az eredmények nyilvánosságra hozatala

A projektben elért eredményeket több fórumon, többféleképpen is publikáltuk. A beszámoló közleményjegyzékében összesen 17 közleményt soroltunk fel, amiből 3 konferenciaabsztrakt, 2 könyvfejezet, 4 folyóiratcikk, 7 konferenciaközlemény és egy szerkesztett mű. Külön kiemeljük a szerkesztett művek jelentőségét – ezek ugyanis a projekt egyik fontos eredményéről tanúskodnak, nevezetesen arról, hogy ez alatt az idő alatt szoros hazai és nemzetközi kapcsolatokat alakítottunk ki a hasonló témával foglalkozó kutatókkal. A projekt elején részt vettünk a *First International Workshop on Computational Linguistics for Uralic Languages* című nemzetközi workshopon, majd a következő évben már a workshop-sorozat második kiadásának társszervezőjeként léptünk fel. Ez az esemény több publikációt is eredményezett: egy folyóiratban közölt konferenciariportot, egy konferenciakiadványt és később egy általunk szerkesztett *Acta Linguistica Academica* különszámot. A workshop-sorozat minden további kiadásán részt vettünk valamilyen formában – programbizottsági tagként, illetve szerzőként. A projekt résztvevői közül többen tagjai lettek az *Association for Computational Linguistics Special Interest Group on Uralic Languages* témacsoportjának.

A finn partnerrel elsősorban az interneten keresztül tartottuk a kapcsolatot, de személyes találkozóra is sor került – jellemzően nemzetközi eseményekre szervezve. Az együttműködésünk egy közös poszterprezentációt eredményezett egy nemzetközi konferencián, továbbá dolgozunk egy közös folyóiratcikken is.

A közleményjegyzékben felsorolt publikációk mellett a projekt eredményeit igyekeztünk hazai és nemzetközi konferenciákon is bemutatni. Az Országos Tudományos Kutatási Alapprogramok és a Finn Akadémia 2012-ben kiírt közös pályázatának a többi nyertesével társprojektekként igyekeztünk együttműködni, ami – többek között – két közösen szervezett eseményt eredményezett. A projektek kezdetekor a *26. Finnugor Szeminárium*on tartottunk egy közös bemutatkozó szekciót, majd a projektek zárásához közeledve szintén közösen szerveztünk egy workshopot *Alkalmazott nyelvészeti kutatások a kisebbségi finnugor nyelvek szolgálatában* címmel. A társkutatókkal és egyéb hazai érdeklődőkkel közösen alakítottunk egy témacsoportot, és létrehoztunk egy levelezőlistát azzal a céllal, hogy a közös jövőbeli terveinket (magyar nyelvű folyóirat-különszám, Tudomány Ünnepe szekció, tematikus szimpózium egy nemzetközi konferencián) könnyebben meg tudjuk valósítani.

Fontosnak tartjuk, hogy az általunk létrehozott erőforrásokat szabadon elérhetővé és használhatóvá tegyük, ezért fejlesztettünk egy weboldalt[14], amin keresztül láthatóvá tudjuk tenni a projekt eredményeit. A weboldalon rövid ismertetést adunk a projektről és a társprojektekről, felsoroljuk a publikációinkat és a konferenciaszerepléseinket, letölthető formában kitesszük a szótárainkat, valamint egy keresőt is fejlesztettünk, amivel mind a 24 általunk előállított szótárban lehet keresni. Az általunk előállított lexikai erőforrások így tehát jelenleg háromféleképpen érhetők el: a weboldalon keresztül, a Wiktionaryben és a BabelNetben. Ezen felül tárgyalásokat folytatunk a Giellatekno üzemeltetőivel is arról, hogy az általunk előállított szótárak bekerüljenek a Giellatekno keretrendszerébe, ami megduplázná az eredményeink láthatóságát, mivel a Giellatekno a finnugor nyelvekkel foglalkozó kutatók körében igen jól ismert, és szabadon felhasználható.

Hivatkozások

Ács, J. (2014). Pivot-based multilingual dictionary building using Wiktionary. In *Proceedings of LREC '14*, Reykjavík, Iceland. ELRA.

[14] `finnotka.nytud.hu`

Ács, J., Pajkossy, K., and Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria. Association for Computational Linguistics.

Ács, J., Pajkossy, K., and Kornai, A. (2017). Digital vitality of Uralic languages. *Acta Linguistica Academica*, 64(3):327–345.

Bakró-Nagy, M., Duray, Zs., Mus, N., Oszkó, B., Sipos, M., Takács, D., and Várnai, Zs. (2015). 'Gold' mining. Exploitation of an etymological database: Uralonet. *Vestnik Ugrovedenia (Bulletin of Ugric studies)*, 1(20):119–125.

Benyeda, I., Koczka, P., Ludányi, Zs., Simon, E., and Váradi, T. (2015). Finnugor nyelvű közösségek nyelvtechnológiai támogatása online tartalmak létrehozásában. In Tanács, A., Varga, V., and Vincze, V., editors, *XI. Magyar Számítógépes Nyelvészeti Konferencia*, pages 133–144, Szeged. SzTE.

Benyeda, I., Koczka, P., and Váradi, T. (2016). Creating seed lexicons for under-resourced languages. In *Proceedings of the GLOBALEX 2016 workshop*, pages 52–56, Portorož. ELRA.

Erdmann, M., Nakayama, K., Hara, T., and Nishio, S. (2009). An Approach for Extracting Bilingual Terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 5(4):31:1–31:17.

Ferenczi, Zs., Mittelholcz, I., and Simon, E. (2018a). Automatic Generation of Wiktionary Entries for Finno-Ugric Minority Languages. In *Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018)*, page 39–50, Helsinki, Finland. Association for Computational Linguistics.

Ferenczi, Zs., Mittelholcz, I., Simon, E., and Váradi, T. (2018b). Evaluation of Dictionary Creating Methods for Finno-Ugric Minority Languages. In *Proceedings of LREC2018*. Közlésre elfogadva.

Fung, P. and Yee, L. Y. (1998). An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 17th International Conference on Computational Linguistics – Volume 1*, COLING '98, page 414–420, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hazem, A. and Morin, E. (2012). ICA for Bilingual Lexicon Extraction from Comparable Corpora. In *The 5th Workshop on Building and Using Comparable Corpora*, pages 126–133, Istanbul, Turkey.

Héja, E. and Takács, D. (2012). Automatically generated online dictionaries. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC2012)*. European Language Resources Association (ELRA).

Kornai, A. (2013). Digital Language Death. *PLoS ONE*, 8(10).

Mohammadi, M. and GhasemAghaee, N. (2010). Building Bilingual Parallel Corpora Based on Wikipedia. In *2010 Second International Conference on Computer Engineering and Applications (ICCEA)*, volume 2, pages 264–268.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing and Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, page 320–322, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon, E., Benyeda, I. Zs., Koczka, P., and Ludányi, Zs. (2015). Automatic creation of bilingual dictionaries for Finno-Ugric languages. In *Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages*, pages 119–131, Tromsø.

Simon, E., Lendvai, P., Németh, G., Olaszy, G., and Vicsi, K. (2012). *A magyar nyelv a digitális korban – The Hungarian Language in the Digital Age*. Georg Rehm and Hans Uszkoreit (Series Editors): META-NET White Paper Series. Springer.

Simon, E. and Mittelholcz, I. (2017). Evaluation of Dictionary Creating Methods for Under-Resourced Languages. In Ekštein, K. and Matoušek, V., editors, *Text, Speech and Dialogue*, volume 10415 of *Lecture Notes in Artificial Intelligence*, pages 246–254, Prague, Czech Republic. Springer International Publishing.

Simon, E., Mittelholcz, I., and Ferenczi, Zs. (2018a). Automatikus szótárépítés kisebbségi finnugor nyelvekre. In Pletl, R. and Kovács, G., editors, *Trans-Linguistica – Multilingualism and Plurilingualism in Europe*. EME-Scientia Publishing House, Cluj-Napoca. Közlésre elfogadva.

Simon, E., Mittelholcz, I., and Ferenczi, Zs. (2018b). Lexikai erőforrások automatikus előállítása kisebbségi finnugor nyelvekre. In Vincze, V., editor, *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*, pages 260–271, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Tamura, A., Watanabe, T., and Sumita, E. (2012). Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, page 24–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tiedemann, J. (2009). News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007*, pages 237–248. John Benjamins, Borovets.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). Parallel corpora for medium density languages. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005*, number 292 in Current Issues in Linguistic Theory, pages 247–258.

Vulić, I. and Moens, M.-F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, page 449–459, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vulić, I. and Moens, M.-F. (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *53rd ACL*, pages 719–725.