Finnugor nyelvű közösségek nyelvtechnológiai támogatása online tartalmak létrehozásában

Benyeda Ivett, Koczka Péter, Ludányi Zsófia, Simon Eszter, Váradi Tamás

MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr u. 33.
e-mail:
{benyeda.ivett,koczka.peter,ludanyi.zsofia,simon.eszter,varadi.tamas}@nytud.mta.hu

Kivonat A cikkben bemutatott folyamatban levő projekt célja, hogy kisebb finnugor nyelvekre állítson elő nyelvi erőforrásokat, amelyekkel revitalizálni lehet ezeket a veszélyeztett nyelvi közösségeket. A projekt során párhuzamos és összevethető korpuszokból kétnyelvű protoszótárakat állítunk elő, melyeket anyanyelvi beszélők fognak ellenőrizni. A különböző nyelvű, egymásnak megfeleltetett szóalakok morfológiai, lexikai, etimológiai információkkal kibővítve kerülnek majd feltöltésre a Wiktionarybe. A projekt során számolnunk kell azzal a nehézséggel, hogy nyelvtechnológiai erőforrások a kisebb finnugor nyelvekre kevéssé állnak rendelkezésre, ezért a szövegfeldolgozás során nyelvfüggetlen gépi tanulási módszereket alkalmazunk. A projekt összes melléktermékét (modellek, korpuszok, szövegfeldolgozó eszközláncok, elemzett szövegek) nyilvánosan elérhetővé tesszük.

Kulcsszavak: párhuzamos korpusz, összevethető korpusz, automatikus szótárgenerálás, finnugor nyelvek, veszélyeztetett nyelvek

1. Bevezetés

Hagyományosan veszélyeztetett nyelvnek azokat a nyelveket szokták nevezni, amelyeknek kevés a beszélője, azok is az idősebb generációba tartoznak, a beszélők száma egyértelműen csökken, és a nyelvhasználat területe határozottan az informális, családi keretek felé tolódik. Kornai [1] 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. Az újfajta szempontrendszer alapján az is fontos kritérium, hogy az egyes nyelvek mennyire vannak jelen a digitális térben: milyen mennyiségben születnek az interneten nyilvánosan hozzáférhető szövegek az adott nyelven. Kornai szerint a digitális nyelvhalál a következőket foglalja magában: funkció- és ezzel együtt presztízsvesztés, és végül a nyelvi kompetencia elvesztése.

A nyelvtechnológia ebben a kontextusban egyfajta támogató technológiaként tud működni: a szabad és nyilvános nyelvhasználatot támogatva, a nyelvi határokat ledöntve segíti a kommunikációt [2]. Nyelvtechnológiai alkalmazások és

erőforrások viszont leginkább a széleskörűen használt nyelvekre léteznek – ezeket nevezi Kornai [1] ún. viruló nyelveknek. 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.

Cikkünkben egy olyan folyamatban lévő projektet mutatunk be, amelynek célja, hogy segítse a veszélyeztetett finnugor nyelvű közösségeket nyelvük felvirágoztatásában azáltal, hogy online tartalmakat hoz létre az adott nyelveken. A projekt során több veszélyeztetett finnugor és néhány széles körben használt viruló nyelvre állítunk elő protoszótárakat, amelyeket lexikai információkkal gazdagítva feltöltünk a Wiktionarybe.

A munkafolyamat első lépése a szöveggyűjtés, valamint párhuzamos és összevethető korpuszok építése (l. 3.1. fejezet). Az alábbi finnugor nyelvekre gyűjtöttünk szövegeket: komi-zürjén, komi-permják, udmurt, mezei és hegyi mari, valamint északi számi. A kis finnugor nyelvek mellett azokra a viruló nyelvekre is kerestünk szövegeket, amelyek a finnugrisztikában fontos szerepet töltenek be, ezek: angol, orosz, finn és magyar.

A párhuzamos és összevethető korpuszok előfeldolgozása automatikusan történik. Tekintve, hogy kifejezetten a szóban forgó kis finnugor nyelvekre fejlesztett tokenizáló és mondatra bontó eszközök nem léteznek, többféle gépi tanulási módszerrel kísérleteztünk. A nyelvi erőforrások hiánya a morfológiai elemzés szintjén több problémát is felvet: felügyelt gépi tanulási módszerek alkalmazása nem lehetséges, mivel a tanításhoz és teszteléshez morfológiailag annotált szövegekre lenne szükség, ezek viszont nem állnak rendelkezésre (l. 3.2. fejezet).

Az összegyűjtött párhuzamos és összevethető szövegeket felhasználva, többféle szótárépítési metódust követve ún. protoszótárakat állítunk elő (l. 4. fejezet), amelyek jelenleg nyelvpáronként néhány száz fordítási jelöltet tartalmaznak. A későbbiekben ezek alapján készülnek majd el azok a szótárak, amelyeket a vizsgált finnugor nyelvek anyanyelvi beszélői fognak kézzel ellenőrizni és javítani. Ezekben a végső szótárakban lesznek azok a szótári elemek, amelyek bizonyos morfológiai információkkal (szófaj, ragozási paradigma), etimológiai, illetve lexikai-szemantikai adatokkal (szinonimák, antonimák) kibővítve kerülnek feltöltésre a Wiktionarybe.

2. Kitekintés

A kétnyelvű szótáraknak nem csupán a gépi fordításban [3] és a nyelvközi információkinyerésben [4] van kritikus szerepük, hanem egyéb nyelvtechnológiai alkalmazásokban is, például a nyelvtanulásban [5], a számítógépes szemantikában, továbbá számos olyan feladatban, ahol megbízható lexikai-szemantikai információra van szükség [6]. Tekintve, hogy a kézzel történő szótárkészítés rendkívül erőforrásigényes, meglehetősen ritkák a szabadon hozzáférhető hagyományos kétnyelvű szótárak. Komplett kétnyelvű szótárak teljesen automatikus módon történő előállítását a jelenlegi technológia nem teszi lehetővé, de a protoszótárak támogatást tudnak nyújtani a lexikográfiai munkához.

A sztenderd szótárépítési módszerek alapjául párhuzamos korpuszok szolgálnak, amelyek az eredeti nyelvű szöveget és annak fordítását tartalmazzák, jellemzően mondatszinten párhuzamosítva. Viszont, ahogy Rapp [7] fogalmaz: mindig kivételes esetnek számít, ha egy adott doménre és adott nyelvpárra elégséges méretű párhuzamos korpusz áll rendelkezésre; általánosnak inkább az tekinthető, ha nincs ilyen. Ilyen korpuszok ugyanis jobbára csupán a legtöbb erőforrással rendelkező nyelvpárokra léteznek. Ez az egyik oka annak, hogy egyre nő az összevethető (nem párhuzamos) korpuszok előállítása iránti érdeklődés.

Az összevethető korpuszokból történő szótárépítési metodológia sztenderd megközelítése kontextusvektorok hasonlóságát méri a két vizsgált nyelvre [8,7]. Ennek 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. Mivel az általunk vizsgált finnugor nyelvekre nem áll rendelkezésre megfelelő méretű korpusz és szótár, alternatív módszerekkel kell kísérleteznünk. Számos újabb módszert alkalmaztak nem párhuzamos korpuszokból történő fordítási párok kinyerésére, például [9,10,11]. Mivel az idézett cikkekben leírt módszerekhez tartozó forráskódok nem publikusak, az eredmények nem reprodukálhatók. Az egyik legújabb trend a nyelvtechnológiában a neurális hálón alapuló vektoros nyelvmodellek használata, amelyet többek között kétnyelvű szótárak előállítására is alkalmaznak (pl. [12]). Ennek a módszernek az általunk vizsgált nyelvpárokra való adaptálását tervezzük elvégezni a projekt során.

3. Párhuzamos és összevethető korpuszok építése

A célkitűzések megvalósításához első lépésként az interneten elérhető párhuzamos és összevethető szövegeket gyűjtöttünk az általunk vizsgált nyelvpárokra. A szótárépítéshez elengedhetetlenül szükséges a gyűjtött szövegek alapszintű nyelvi feldolgozása (tokenizálás, mondatra bontás, lemmatizálás, morfológiai elemzés és egyértelműsítés). Ebben a fejezetben a korpuszépítési munkafolyamat lépéseit ismertetjük.

3.1. Szöveggyűjtés

McEnery és Xiao [13] definícióját követve akkor beszélhetünk párhuzamos korpuszról, ha a korpuszt felépítő szövegek egy az egyben egymás fordításai. Ha a korpusz különböző nyelvű részei nem pontos fordításai egymásnak, de a mintavétel módját tekintve megegyeznek, akkor összevethető korpuszról beszélünk. Időnként azonban nem teljesen egyértelmű, hogy egy többnyelvű szöveghalmaz párhuzamos vagy inkább összevethető korpuszként kezelendő. Szigorúan véve csak a biblia- és regényfordítások, a szoftverdokumentációk és az olyan hivatalos dokumentumok, mint például az Egyetemes Emberi Jogok Nyilatkozata, tekinthetők valódi párhuzamos szövegeknek. A Wikipédia-szócikkek többé-kevésbé

egymás fordításai, de a szócikkek különböző nyelvbeli megfelelői között igen jelentős különbségek lehetnek, így ezek nem tekinthetők párhuzamos szövegeknek, de összevethető korpuszok építésére jól hasznosíthatók.

Párhuzamos korpuszok. A Bibliának párhuzamos szövegként történő felhasználása régi hagyománnyal bír a szótárépítésben [14], így mi is alkalmaztuk ezt a módszert. A Parallel Bible Corpus [15], a Bible.is és a The Unbound Bible oldalakról letöltöttük az Újszövetség fordításait a szóban forgó nyelvekre. Hogy a készülő szótárak ne tartalmazzanak archaikus és kihalt szavakat, mindig a legújabb bibliafordítást választottuk. A fordítások versszinten párhuzamosítva vannak, így a szövegek további feldolgozása könnyűszerrel megoldható. Nehézséget jelent azonban, hogy bizonyos nyelvekre (udmurt, hegyi mari, komi-permják) nem találtunk elektronikus szöveges formátumban elérhető bibliafordításokat.

A Biblián kívül további párhuzamos korpuszként használható az OPUS korpusz [16], amelyben találtunk északi számi szoftverdokumentációt párhuzamosítva mind a négy viruló nyelvi megfelelőjével. Párhuzamos korpuszok forrásaként használhatók továbbá Finnország és Norvégia egyes hivatalosan kétnyelvű régióinak weboldalai is. Olyan párhuzamos szövegeket, ahol az egyik nyelv komi-permják, hegyi mari, illetve udmurt, sajnos nem találtunk.

Összevethető korpuszok. Az összevethető korpuszok előállításához az egyik leggyakrabban használt forrás a Wikipédia. A munka első fázisaként minden általunk vizsgált nyelvre letöltöttük a Wikipédia dump fájlokat. Ezt követően a nyelvközi linkek segítségével összepárosítottuk az azonos témájú, különböző nyelveken íródott cikkeket. A szövegek kinyeréséhez a Wikipedia Extractort[1] használtuk, annyi módosítással, hogy a cikkszövegek mellett megtartottunk néhány egyéb metaadatot is (szövegben előforduló nyelvközi linkek, Wikidata-azonosítók), amelyek a további feldolgozást segítik. A Wikidata a Wikipédia testvérprojektje: egy ingyenes, közösség által szerkesztett, többnyelvű tudásbázis, ahol a nyelvközi linkek által összekapcsolt Wikipédia-címszavak egy és ugyanazon entitáshoz tartoznak, egyetlen Wikidata-azonosítóval. Ezek az azonosítók alkalmasak arra, hogy megtaláljuk a szövegekben lévő azonos névelemeket – a cikk nyelvétől függetlenül –, valamint horgonyként használva segítséget nyújthatnak a szövegek párhuzamosításában is.

Míg a nagyméretű, aktív digitális közösség által szerkesztett és karbantartott Wikipédia-cikkek elég terjedelmesek, a kis nyelvi közösséggel rendelkező Wikipédiák esetében a cikkek mennyisége és terjedelme jóval kisebb. Ebből kifolyólag az egymásnak megfelelő különböző nyelvű cikkek hossza általában meglehetősen eltérő. Feltételezve, hogy a cikkek első, definíciós része minden nyelven nagyjából megfelel egymásnak, minden cikkpárnál a szöveg első x mondatát tartottuk meg, ahol x egyenlő a finnugor nyelvű cikk mondatainak számával.

A sztenderd megközelítés szerint (pl. [8]) azok a szövegek is összevethető korpuszokként kezelhetők, amelyek két vagy több nyelvű újságcikkeket, híreket

[1] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

tartalmaznak ugyanabból az időintervallumból és ugyanarról a helyről. Ez utóbbi konstelláció miatt feltételezhetjük, hogy a cikkek ugyanazokról a helyi vagy fontosabb globális történésekről számolnak be, így ha nem is egymás fordításai, témában nagyon közel állnak egymáshoz. Erre alapozva gyűjtöttünk cikkeket finnországi online újságok honlapjairól északi számi–{finn, angol, orosz} nyelvpárokra.

További módszer összevethető korpuszok építésére az azonos téma köré szerveződő alkorpuszok felhasználása, vagyis olyan egynyelvű szövegek letöltése különböző nyelveken, amelyek azonos tárgykörhöz tartoznak [8]. E korpuszok létrehozásához olyan szövegeket töltöttünk le északi számi és angol nyelven, amelyek a számi kultúráról, oktatásról és társadalomról szólnak.

Egynyelvű szövegek. Az előbbiek mellett minden kis finnugor nyelvre létrehoztunk egynyelvű korpuszokat is. Míg a párhuzamos és összevethető korpuszokat szótárépítésre használjuk, az egynyelvű korpuszok tanítóanyagként funkcionálnak a tokenizáló és mondatra bontó alkalmazások számára. Az egynyelvű korpuszokat felépítő anyagokat különféle weboldalakról töltöttük le, így ezek témában igen változatosak (pl. irodalmi szövegek, hírek, személyes blogok, hivatalos szövegek).

Az 1. táblázat a párhuzamos, az összevethető és az egynyelvű korpuszok tokenszámát mutatja. Az egynyelvű szövegek tokenszámába a párhuzamos és az összevethető korpuszok adott nyelvű részei is bele vannak számolva. Az összevethető korpuszok számadatai a csökkentett méretű Wikipédia-cikkek szövegeire vonatkozóan értelmezendők, vagyis a cikkeknek kizárólag az első x mondatát tartalmazzák (lásd feljebb). Az összevethető korpuszok közül az időintervallum-alapúak évenkénti bontásban lettek számolva, vagyis nem számoltuk bele azokat a szövegeket, amelyeknek nem volt ugyanazon évből származó másik nyelvű megfelelője. A táblázat adatai a szöveggyűjtés jelenlegi állapotát mutatják; a számok a projekt előrehaladtával természetesen változnak.

3.2. Szövegfeldolgozás

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. Sajnos a cirill betűs finnugor nyelvekre nem találtunk tokenizáló és mondatra bontó eszközöket. Az egyetlen kis finnugor nyelv, amely nyelvtechnológiai eszközökkel kellően támogatott, az a latin ábécés északi számi. Ez nem okozhat különösebb meglepetést, hiszen az északi számi rendelkezik a legjelentősebb mértékű online forrásokkal, beleértve ebbe a Tromsői Egyetemen fejlesztett eszközöket[2] és az online elérhető tartalmakat.

[2] http://giellatekno.uit.no/cgi/index.sme.eng.html

1. táblázat. Az egynyelvű, párhuzamos és összevethető korpuszok tokenszámai. A táblázatban szereplő ISO 639-3 nyelvkódok: sme – északi számi, kpv – komi-zürjén, koi – komi-permják, mhr – mezei mari, mrj – hegyi mari, udm – udmurt; eng – angol, fin – finn, rus – orosz, hun – magyar.

nyelv	**egynyelvű**	**nyelvpár**	**párhuzamos**		**összevethető**	
			L1	**L2**	**L1**	**L2**
sme	1.364.254	sme-eng	691.260	724.750	253.930	1.754.968
		sme-fin	245.440	273.973	239.651	5.259.591
		sme-rus	173.179	220.790	212.332	233.748
		sme-hun	171.668	224.014	86.244	106.391
kpv	480.609	kpv-eng	121.108	174.742	89.580	183.602
		kpv-fin	121.120	133.715	88.507	80.797
		kpv-rus	117.903	125.085	108.013	141.369
		kpv-hun	121.319	134.344	68.179	74.274
koi	719.325	koi-eng	0	0	257.871	194.784
		koi-fin	0	0	137.578	77.696
		koi-rus	0	0	188.334	139.976
		koi-hun	0	0	95.120	64.794
mhr	1.335.457	mhr-eng	128.316	175.075	121.588	250.583
		mhr-fin	128.328	133.965	118.120	115.028
		mhr-rus	109.449	109.818	158.977	215.724
		mhr-hun	128.565	134.618	106.813	121.453
mrj	366.964	mrj-eng	0	0	137.088	306.465
		mrj-fin	0	0	85.134	93.622
		mrj-rus	0	0	124.289	187.687
		mrj-hun	0	0	77.855	90.168
		mrj-hun	0	0	77.855	90.168
udm	584.113	udm-eng	0	0	67.306	135.450
		udm-fin	0	0	56.222	49.961
		udm-rus	0	0	80.800	129.293
		udm-hun	0	0	41.883	48.736

Előfeldolgozás. Ahogy a 3.1. fejezetben szó volt róla, viszonylag nagy mennyiségű egynyelvű szöveget gyűjtöttünk minden nyelvre. Nehézséget jelent azonban, hogy az egyes szövegek több helyütt tartalmaznak nem odavaló szövegrészeket.

Az első nehézség, hogy a cirill betűs finnugor nyelvek az ábécé módosított verzióit használják, amelyekben sok a különféle diakritikus jelekkel ellátott cirill karakter. Ezek a speciális karakterek gyakran az alapkarakter és a diakritikus jel kombinációjából állnak elő, amelyek így a későbbi lépésekben használt eszközök számára külön karakterekként értelmeződnek. Ennek elkerülése érdekében a további szövegfeldolgozás előtt karakternormalizálást kell végezni minden forráson.

A második, szintén jelentős probléma a cirillel írt nyelvek esetében, hogy az egymással közeli rokonságban álló nyelvek (komi-permják és zürjén, mezei és hegyi mari) gyakran egyszerre is megjelenhetnek egy dokumentumon belül, ezért az ilyen részeket el kell különítenünk egymástól. A nyelvek megkülönböztetésére a Blacklist Classifiert[3] használtuk, amely 97,47%-os pontossággal szűri a komi-zürjén és komi-permják, 96,77%-os pontossággal pedig a mezei és hegyi mari nyelveket.

A harmadik probléma, hogy bizonyos finnugor nyelvű személyes blogok angol vagy orosz nyelvű blogszolgáltatókon találhatók, így az egyébként egynyelvű blogbejegyzések nyelvileg kevertek, mivel számos hasznosítható információ, a dátumok és a weblap bizonyos elemei nem a kívánt finnugor nyelven szerepelnek a szövegben. Az idegen nyelvű részek kiszűrésére egy trigramstatisztikát és Katz–Backoff-simítást alkalmazó nyelvfelismerő szkriptet, a Langid-t[4] használtuk. A szükséges modellek kézzel válogatott szövegek felhasználásával készültek. A dátumokat minden esetben megtartottuk, mivel ez alapján az információ alapján tudjuk előállítani az időintervallum-alapú összevethető korpuszokat.

Mondatra bontás, tokenizálás. A mondatra bontáshoz és tokenizáláshoz az Apache OpenNLP[5] mondatra bontó és tokenizáló moduljait használtuk. Ahogy már említettük, az északi számi igen jól támogatott NLP-eszközöket illetően, ezért csak a cirill ábécét használó finnugor nyelvekre építettünk modelleket. Az OpenNLP-nek mind a tokenizáló, mind a mondatra bontó eszköze 98% feletti F-mértékkel teljesített. Ez a teljesítmény részben annak köszönhető, hogy a mondatra bontó rövidítésszótár használatát is lehetővé teszi, így ennek segítségével elkerülhető az, hogy az eszköz tévesen mondathatárként ismerje fel a rövidítések utáni pontot. A mondatra bontásban bevett szokás a rövidítéslisták használata, bár nyilvánvalóan nem lehetséges egy teljeskörű lista létrehozása, különösen a kutatásunkban releváns finnugor nyelvek esetében. Az általunk használt rövidítésszótár a Wiktionary orosz rövidítéslistáján alapul; ezt a későbbiekben kibővítjük a kis finnugor nyelvekben előforduló rövidítésekkel. Mindazonáltal az orosz rövidítéseket tartalmazó szótár használata jelenleg elégségesnek bizonyult, hiszen a cirillel írt finnugor nyelvekben használt rövidítések gyakran azonosak az orosz rövidítésekkel.

[3] https://bitbucket.org/tiedemann/blacklist-classifier/wiki/Home
[4] https://github.com/juditacs/langid
[5] https://opennlp.apache.org/

Morfológiai elemzés és egyértelműsítés. Online hozzáférhető morfológiai elemző és/vagy egyértelműsítő elérhető északi számi[6], udmurt és komi-zürjén[7], valamint hegyi mari[8] nyelvekre. Legjobb tudomásunk szerint azonban mezei mari és komi-permják nyelvekre nem létezik morfológiai elemző. Ezen a feldolgozási szinten az erőforrások hiánya még komolyabb problémát okoz, hiszen ezekre a nyelvekre még morfológiai annotációval ellátott szövegek sincsenek, amiken tanítani lehetne egy felügyelt gépi tanuló rendszert.

A morfológiai elemzővel nem rendelkező nyelvek esetében több lehetőség közül választhatunk. Az egyik opció, ha egy félig felügyelt vagy felügyelet nélküli morfológiai szegmentáló eszközt használunk (pl. Morfessor[9]), annak működését igényeink szerint kibővítve. Végeztünk ilyen irányú kísérleteket a Morfessorral: egy udmurt szólistán betanítottuk, és a szegmentált kimenetet összehasonlítottuk egy udmurt morfológiai elemző [17] kimenetével. A biztató eredmények ellenére azonban jelentős munkát igényelne olyan további eszközök fejlesztése, amelyek lemmákat és morfológiai címkéket adnának a Morfessor kimenetéhez.

Egy másik lehetőség az, hogy a közeli rokonságban álló nyelvekre alkalmazunk már létező eszközöket. A legegyszerűbb megoldás az, ha egy eszközt közvetlenül a rokon nyelvre alkalmazunk, vagyis pl. a komi-zürjénre kifejlesztett modellt közvetlenül használjuk a komi-permják adatokon. Reményeink szerint a komi-zürjén morfológiai jegyek közvetlenül átvihetők az azonos szöveg komi-permják változatára, valamint ugyanígy járnánk el a hegyi és mezei mari nyelvek esetében. Mivel általánosságban véve a nyelvek nagy részére nem áll rendelkezésre elégséges mennyiségű tanítóanyag, a közeli rokonságban álló nyelvek közötti annotációátvitelre épülő különböző kísérletek az utóbbi időben kitüntetett szerepet élveznek az NLP-kutatásokban (pl. [18,19]). Terveink között szerepel annak további vizsgálata, hogy hogyan lehet átültetni az annotációkat egy nyelvtechnológiailag jobban támogatott nyelv kevésbé támogatott közeli rokon nyelvére.

4. Protoszótárak építése

A jelenlegi módszerek nem teszik lehetővé kétnyelvű szótárak teljesen automatikusan történő létrehozását, de kivitelezhető bizonyos lexikai erőforrások, ún. protoszótárak gépi előállítása, amelyek nagy segítséget jelenthetnek a szótárépítési munkálatokban. A hagyományos, manuálisan készített szótáraknál a protoszótárak általában nagyobb méretűek, és a lexikai elemek nagyobb lefedettségét biztosítják, ugyanakkor jóval több nem megfelelő fordítási jelöltet tartalmaznak. A protoszótárak méretének kiválasztása ezért nagyban függ a későbbi felhasználás módjától.

Többféle szótárépítési módszerrel kísérleteztünk, melyeket az alábbiakban ismertetünk. Mindegyik módszerrel több száz fordítási jelöltet tartalmazó protoszótárat hoztunk létre majdnem minden nyelvpárra. Ezek a szótárfájlok a

[6] http://giellatekno.uit.no/cgi/index.sme.eng.html
[7] http://www.morphologic.hu/urali/
[8] http://www.univie.ac.at/maridict/site-2014/morph.php
[9] http://morfessor.readthedocs.org/en/latest/general.html#techrep

későbbi végleges szótárak kiindulási pontjaként fognak szolgálni, melyekbe csak a legvalószínűbb fordítási párok fognak bekerülni. A fordítások kézi ellenőrzését anyanyelvi beszélők végzik majd a projekt utolsó szakaszában.

4.1. Létező szótárak felhasználása

Online szótárakat minden finnugor nyelv legalább egy nyelvpárjára találtunk. Ezek többségében online használhatók, néhány azonban letölthető változatban is hozzáférhető volt. A HTML-fájlok feldolgozásával megtörtént a szópárok kinyerése, aminek eredményeképpen kétnyelvű szótárak jöttek létre néhány nyelvpárra, átlagosan néhány száz szópár méretben. Ezek a szótárak az automatikus szótárépítési munkák során segítséget nyújthatnak egyrészt a párhuzamos vagy összevethető cikkpárok mondat- és frázisszintű illesztésénél, másrészt a szótárgeneráló algoritmusok számára kiinduló, ún. magszótárakként szolgálhatnak.

4.2. Wikipédia-címszavak

A Wikipédia nemcsak a legnagyobb, szabadon elérhető adatbázis, amely összevethető szövegeket tartalmaz, de többféle módon is felhasználható kétnyelvű szótárak létrehozására. Erdmann et al. [20] címszavakból készített kétnyelvű szótárakat használt a cikkek szövegeiből történő további fordítási párok kinyeréséhez. Mohammadi és Quasim Aghaee [21] az angol és a perzsa Wikipédiából párhuzamos mondatpárokat nyert ki, szintén Wikipédia-címszavakból épített szótárakat felhasználva. Ezt a módszert követve kétnyelvű szótárakat hoztunk létre Wikipédia-címpárokból a nyelvközi linkek segítségével, amely nyelvpáronként további néhány száz elemű szótárat eredményezett.

4.3. Wiktionaryre épülő módszerek

A Wikipédia mellett a Wiktionary egy másik, szintén nyílt, közösség által szerkesztett tudásbázis, amely kiváló forrásul szolgálhat kétnyelvű szótárak létrehozásához. Bár a Wiktionary elsősorban emberi felhasználásra készült, a benne található adatok kinyerése bizonyos fokig automatizálható. Ács et al. [22] minden Wictionary-cikkhez tartozó fordítási elemet kinyert a cikkekben található fordítási táblákból. Mivel az általuk fejlesztett Wikt2dict eszköz[10] szabadon elérhető, fel tudtuk használni a projektünkben szerepet kapó nyelvpárokra. Az angol, finn, orosz és magyar Wiktionary oldalak parszolásával szinte minden szóban forgó nyelvpárra számos fordítási szópárt sikerült kinyernünk.

Ács [23] a szópárok halmazát újabbakkal bővítette, oly módon, hogy új kapcsolatokat hozott létre a már meglévő fordítási párokból egy ún. háromszögelési módszerrel. A háromszögelés azon a feltételezésen alapul, hogy két elem nagy valószínűséggel fordításpár abban az esetben, ha mindkettő egy harmadik nyelv szavának fordításpárja. A Wikt2dict háromszögelési technikájával szótárainkat további néhány száz elemmel tudtuk bővíteni.

[10] https://github.com/juditacs/wikt2dict

4.4. Hundict

A Hundict[11] egy kísérleti projekt kétnyelvű szótárak párhuzamos szövegekből való előállítására. A program az egymásnak megfeleltetett szövegekből nyeri ki a gyakran együtt előforduló szópárokat a Sørensen–Dice-együttható alapján. Az eszköz hatékonysága növelhető gold standard szótárak és stopszólisták hozzáadásával. Végeztünk néhány kísérletet bibliafordításokkal északi számi–finn és komi-zürjén–angol nyelvpárokra, amelyek eredményképpen olyan fordítási párokat kaptunk, ahol a szópárok mellett a fordítások konfidenciaértéke is szerepel. A rendszer több olyan paramétert is tartalmaz, melyek további finomhangolása szükséges az elérhető legjobb eredmény érdekében. Terveink közt szerepel ezek kimérése és további nyelvpárokra való felhasználása is, bár az eszköz lemmatizált szövegeket kíván bemenetként, így előbb a párhuzamos korpuszok lemmatizált változatát kell elkészítenünk.

5. Összegzés, további feladatok

Cikkünkben egy olyan folyamatban lévő projektet mutattunk be, amelynek forrásául a webről letöltött párhuzamos és összevethető korpuszok szolgálnak. A projekt fő célja, hogy olyan protoszótárakat hozzunk létre automatikus módszerekkel, amelyeknek egyik nyelve az alábbi kis finnugor nyelvek egyike: komi-zürjén, komi-permják, mezei és hegyi mari, udmurt, valamint északi számi. Ezekre a nyelvekre kevés nyelvtechnológiai eszköz készült, sőt bizonyos nyelvek esetében még digitális szöveges tartalmak is csupán igen kis számban lelhetők fel, ebből kifolyólag mind a szöveggyűjtés, mind a szövegfeldolgozás nagy kihívásokat jelent. A szövegfeldolgozás alsóbb szintjein problémát jelent az, hogy kifejezetten a szóban forgó kis finnugor nyelvekre nem létezik tokenizáló és mondatra bontó eszköz. A morfológiai elemzés szintjén további nehézségekkel kell számolnunk: nem alkalmazhatók felügyelt gépi tanulási módszerek, mivel nem áll rendelkezésünkre tanításhoz és teszteléshez szükséges morfológiai annotációval ellátott szöveg.

A leírt nehézségek ellenére kellő méretű egy- és többnyelvű szöveget gyűjtöttünk az általunk vizsgált nyelvpárokra, melyeket felhasználva nyelvfüggetlen módszerekkel protoszótárakat állítottunk elő. A létrehozott szótárakat bizonyos morfológiai, etimológiai, szemantikai információkkal és többnyelvű fordítási megfelelőikkel kibővítve feltöltjük a Wiktionarybe. Mind a szótárakat, mind a nyelvészeti információkat a lehetőségekhez mérten automatikusan állítjuk elő. Természetesen nem nélkülözhető a kézi kiértékelés és javítás, ezt anyanyelvi beszélők fogják végezni a projekt utolsó szakaszában.

A Wiktionary rendszerét felhasználva a szótári elemek a Wiktionary különböző nyelvű változataiban összekapcsolhatók. Ez lehetővé teszi, hogy a közösség gazdag lexikai anyagokhoz férjen hozzá, mindemellett olyan új adatok is elérhetők lesznek a lexikai elemekhez, mint például az etimológiai adatok vagy a fordítási megfelelők. A Wiktionary sajátos markup formátumot használ, amit

[11] https://github.com/zseder/hundict

mi XML-formátumra alakítunk a további feldolgozás érdekében. A jogi kérdések tisztázása után az összes létrehozott anyagot: a korpuszokat, a szótárakat, illetve a nyelvmodelleket publikusan elérhetővé tesszük.

Köszönetnyilvánítás

A projektet az Országos Tudományos Kutatási Alapprogram támogatja (szerződésszám: 107885).

Hivatkozások

1. Kornai, A.: Digital Language Death. PLoS ONE **8**(10) (2013)
2. Simon, E., Lendvai, P., Németh, G., Olaszy, G., Vicsi, K.: 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 (2012)
3. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1) (2003) 19–51
4. Grefenstette, G.: The Problem of Cross-Language Information Retrieval. In Grefenstette, G., ed.: Cross-Language Information Retrieval. Kluwer Academic Publishers (1998) 1–9
5. Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J.B., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., Volodina, E.: Corpus-Based Vocabulary lists for Language Learners for Nine Languages. Language Resources and Evaluation **48**(1) (2013) 121–163
6. Zesch, T., Müller, C., Gurevych, I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08), Marrakech, Morocco, ELRA (2008)
7. Rapp, R.: Identifying word translations in non-parallel texts. In: Proceedings of the 33rd annual meeting on Association for Computational Linguistics. ACL '95, Stroudsburg, PA, USA, Association for Computational Linguistics (1995) 320–322
8. Fung, P., Yee, L.Y.: 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, Stroudsburg, PA, USA, Association for Computational Linguistics (1998) 414–420
9. Hazem, A., Morin, E.: ICA for Bilingual Lexicon Extraction from Comparable Corpora. In: The 5th Workshop on Building and Using Comparable Corpora, Istanbul, Turkey (2012) 126–133
10. Tamura, A., Watanabe, T., Sumita, E.: 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, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 24–36
11. Vulić, I., Moens, M.F.: 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, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 449–459

12. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: Distributed Word Representations for Multilingual NLP. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, Association for Computational Linguistics (2013) 183–192
13. McEnery, A., Xiao, R.: Parallel and comparable corpora: What are they up to? In James, G., Anderman, G., eds.: Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters (2007)
14. Resnik, P., Olsen, M.B., Diab, M.: The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. Computers and the Humanities **33**(1–2) (1999) 129–153
15. Mayer, T., Cysouw, M.: Creating a massively parallel Bible corpus. In: Proceedings of LREC '14, Reykjavik, Iceland, ELRA (2014)
16. Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, ELRA (2012)
17. Novák, A.: Morphological Tools for Six Small Uralic Languages. In: Proceedings of LREC '06, ELRA (2006)
18. Scherrer, Y., Sagot, B.: A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, ELRA (2014) 502–508
19. Ingason, A.K., Loftsson, H., Rögnvaldsson, E., Sigurdsson, E.F., Wallenberg, J.C.: Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian. In: Proceedings of LREC'14, Reykjavik, Iceland, ELRA (2014) 91–95
20. Erdmann, M., Nakayama, K., Hara, T., Nishio, S.: An Approach for Extracting Bilingual Terminology from Wikipedia. ACM Transactions on Multimedia Computing, Communications, and Applications **5**(4) (2009) 31:1–31:17
21. Mohammadi, M., GhasemAghaee, N.: Building Bilingual Parallel Corpora Based on Wikipedia. In: 2010 Second International Conference on Computer Engineering and Applications (ICCEA). Volume 2. (2010) 264–268
22. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, Association for Computational Linguistics (2013) 52–58
23. Ács, J.: Pivot-based multilingual dictionary building using Wiktionary. In: Proceedings of LREC '14, Reykjavik, Iceland, ELRA (2014)