Lexikai erőforrások automatikus előállítása kisebbségi finnugor nyelvekre

Simon Eszter, Mittelholcz Iván, Ferenczi Zsanett

MTA Nyelvtudományi Intézet
{VEZETÉKNÉV.KERESZTNÉV}@nytud.mta.hu

Kivonat A cikkben bemutatott projekt célja, hogy kisebbségi finnugor nyelvek számára nyelvi erőforrásokat állítson elő, melyek segítik ezeket a veszélyeztetett nyelvi közösségeket a revitalizálási folyamatokban. A projekt során kétnyelvű protoszótárakat állítottunk elő, melyeket anyanyelvi beszélők és nyelvész szakértők ellenőriztek. Az ellenőrzött fordítási párok különféle nyelvi információkkal kiegészítve kerülnek feltöltésre a Wiktionarybe. A cikk bemutatja a kétnyelvű szótárak automatikus előállításának és kézi validálásának menetét, valamint azt a munkafolyamatot, amivel a Wiktionary-szócikkeket állítottuk elő teljesen automatikusan.
Kulcsszavak: kétnyelvű szótárak, lexikai erőforrások, finnugor nyelvek, kisebbségi nyelvek, számítógépes nyelvészet, Wiktionary

1. Bevezetés

Cikkünkben egy olyan projektet mutatunk be, amelynek célja, hogy segítse a veszélyeztetett finnugor nyelvű közösségeket a digitális revitalizációban online tartalmak létrehozásával. A projekt során kétnyelvű szótárakat állítottunk elő, melyekkel kisebbségi finnugor nyelvek revitalizációját próbáljuk támogatni úgy, hogy az addig csak szegényes digitális tartalommal rendelkező nyelvek számára újabb szópárokkal gazdagítjuk az interneten fellelhető fordítási párok számát. A feltöltendő szópárokat nyelvi információkkal bővítjük ki, és a kész szócikkeket elérhetővé tesszük a Wiktionary keretein belül.

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. Emellett olyan új adatok is elérhetők lesznek a lexikai elemekhez, mint a szófaji információ vagy a fordítási megfelelők. Ezzel támogatni kívánjuk a veszélyeztetett finnugor nyelvű közösségeket a digitális revitalizációban, amivel – reményeink szerint – hozzájárulunk a nyelvi sokszínűség fenntartásához. Szabadon elérhető online többnyelvű lexikai erőforrás a sok beszélővel rendelkező nyelvekre is kevés van – kivétel ez alól a BabelNet [1] és a szabadon elérhetővé tett többnyelvű wordnetek, mint például a MultiWordNet [2] –, vagyis a megbízható lexikai erőforrások előállítása minden nyelvre kiemelten hasznos.

A projekt során hat kisebbségi finnugor nyelvvel dolgozunk forrásnyelvként: komi-permják, komi-zürjén, udmurt, mezei mari, hegyi mari és északi számi. A

fordítások célnyelve négy olyan, sok beszélővel rendelkező nyelv, melyek ezen nyelvközösségek szempontjából fontos szerepet játszanak: az angol, a finn, a magyar és az orosz. Egy forrásnyelv négy célnyelvvel alkothat párt, így összesen 24 nyelvpárral dolgozunk.

Az Expanded Graded Intergenerational Disruption Scale (EGIDS) [3] egy olyan skála, amely meghatározza a világ nyelveinek helyzetét, és segítségével osztályozható a nyelvek vitalitásának mértéke. A legmagasabb szint a 0., az ide tartozó nyelveket nemzetközi szinten, számos funkcióban használják, míg a 10. szinten levők kihalt nyelveknek számítanak. Az 1. táblázat összefoglalja az említett finnugor nyelvek jellemzőit. Minden nyelv mellett szerepel az ISO-639-3 nyelvkódja (a további táblázatokban is ezeket használjuk), az EGIDS-szintje, a beszélőinek a száma, az ország(ok), ahol beszélik és az írásrendszer, amit használ. Érdemes megfigyelni, hogy a legkevésbé veszélyeztetett nyelv az északi számi, annak ellenére, hogy a beszélőinek a száma a legkevesebb a felsorolt nyelvek közül. Ezek a számok az északi számi nyelvet célzó revitalizációs törekvések sikerességéről tanúskodnak.

nyelv	**ISO**	**EGIDS**	**népesség**	**terület**	**írás**
északi számi	sme	2	26.000	Norvégia, Finnország, Svédország	latin
mezei mari	mhr	4	470.000	Oroszország	cirill
hegyi mari	mrj	5	30.000	Oroszország	cirill
komi-zürjén	kpv	5	156.000	Oroszország	cirill
komi-permják	koi	5	63.000	Oroszország	cirill
udmurt	udm	5	340.000	Oroszország	cirill

1. táblázat. A kisebbségi finnugor nyelvek jellemzőinek összefoglalója.

2. Automatikus szótárépítés

2.1. Kitekintés

A kétnyelvű szótáraknak nem csak a nyelvtanulásban és a lexikográfiában van fontos szerepük, hanem olyan nyelvtechnológiai alkalmazásokban is, mint a gépi fordítás [4] és a nyelvközi információ-visszakeresés [5]. 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 sztenderd megközelítése párhuzamos vagy összevethető korpuszokból történő kontextushasonlóság-számításon alapul [6]. Az elmúlt években a forrás- és célnyelvi szavakat reprezentáló vektorokat jellemzően szóbeágyazást alkalmazó módszerekkel nyerik ki [7]. Ezeknek a módszereknek az a hátránya, hogy nagy mennyiségű szöveget igényelnek. Viszont kivételes esetnek számít, ha egy adott nyelvpárra elég nagy párhuzamos vagy összevethető korpusz áll rendelkezésre; általánosnak inkább az tekinthető, ha nincs ilyen. Mivel az általunk vizsgált finnugor nyelvekre nincs kellően nagy korpusz, alternatív módszerekkel kísérleteztünk.

2.2. Az alkalmazott módszerek

A fent leírt okokból a protoszótárak előállításához két, közösség által épített nyelvi erőforrást használtunk fel, a Wikipédiát és a Wiktionaryt.

A Wikipédia[1] többféle módon is felhasználható kétnyelvű szótárak létrehozására. Mi Erdmann et al. [8] és Mohammadi és GhasemAghaee [9] módszerét követve kétnyelvű szótárakat hoztunk létre Wikipédia-címszópárokból a nyelvközi linkek segítségével.

A Wikipédia mellett a Wiktionary[2] egy másik, szintén nyílt, közösség által szerkesztett tudásbázis, amely forrásul szolgálhat kétnyelvű szótárak létrehozásához. A Wiktionary egy olyan többnyelvű szótár, amelynek célja, hogy minden nyelv minden szavát tartalmazza. Sok nyelven elérhető, és a definíciók, leírások mindig az adott Wiktionary nyelvén szerepelnek. Például a Wiktionary magyar kiadását Wikiszótárnak hívják, és a magyar szavak mellett tartalmaz más nyelvű szavakat is, de a hozzájuk tartozó definíciók és nyelvi információk magyarul szerepelnek. Nyelvenként eltérő, hogy mi az a minimális információ, amit tartalmaznia kell egy szócikknek, de maga a címszó, valamint az, hogy milyen nyelvű és milyen szófajú, kötelező információ minden Wiktionary-kiadás esetében. Nem kötelező elem, de bizonyos nyelvű Wiktionarykben szerepel egy fordítási tábla is, ami a címszó különböző nyelvű fordításait tartalmazza; szerepelhetnek továbbá a címszó IPA-átirata, ragozási paradigmája, szinonimái és egyéb lexikai információk is az egyes szócikkekben.

Bár a Wiktionary elsősorban emberi felhasználásra készült, a benne található adatok kinyerése bizonyos fokig automatizálható. Ács et al. [10] minden címszóhoz tartozó fordítási megfelelőt kinyert a szócikkekben található fordítási táblákból. Az általuk fejlesztett `Wikt2dict`[3] eszközzel feldolgoztuk az angol, finn, orosz és magyar Wiktionary-oldalakat, így szinte minden szóban forgó nyelvpárra sikerült fordítási párokat kinyernünk.

Ács [11] 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

[1] `https://www.wikipedia.org/`
[2] `https://www.wiktionary.org/`
[3] `https://github.com/juditacs/wikt2dict`

`Wikt2dict` háromszögelési technikájával protoszótárainkat tovább tudtuk bővíteni.

2.3. Kiértékelés

Első lépésként 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. Az instrukciók, melyek szerint dolgoztak, az alábbiak: a forrás- (S) és célnyelvi (T) szónak is az adott nyelven létező szónak kell lennie, szótári alakban kell állnia, és a két szónak egymás fordításának kell lennie. Ha a forrásnyelvi szó nem létező szó, a szópárt hibásnak kell jelölni. Ha a forrásnyelvi szó létező szó, de nem szótári alakban áll, meg kell adni a helyes szótári alakot. Ha a célnyelvi szó jó fordítása a forrásnyelvi szónak, de nem szótári alakban áll, meg kell adni a helyes szótári alakot. Ha a célnyelvi szó nem jó fordítás, új fordítást kell megadni.

A kiértékelés során bevezettünk olyan kategóriákat, melyek azt jelzik, hogy egy adott szópár megfelel-e ezeknek az instrukcióknak, vagy sem. A kategóriák a következők:

ok-ok: az S és T szavak létező, szótári alakban álló szavak, és egymás fordításai;
ok-nd: az S és T szavak létező szavak, egymás fordításai, de a T szó nem szótári alakban szerepel;
nd-ok: az S és T szavak létező szavak, egymás fordításai, de az S szó nem szótári alakban szerepel;
nd-nd: az S és T szavak létező szavak, egymás fordításai, de sem az S, sem a T szó nem szótári alakban szerepel;
ok-wr: az S szó létező, szótári alakban szereplő szó, de a T szó vagy nem létező szó, vagy nem helyes fordítása az S szónak;
nd-wr: az S szó létező, de nem szótári alakban szereplő szó, a T szó pedig vagy nem létező szó, vagy nem helyes fordítása az S szónak;
wr-xx: az S szó az adott nyelven nem létező szó.

Az automatikusan létrehozott protoszótárak kézi kiértékelése és javítása több célt is szolgál. Egyrészt lehetőséget ad az általunk használt szótárépítési módszerek összehasonlítására – ezt ismertetjük ebben a fejezetben. Másrészt megadja azoknak a szópároknak a számát, amelyeket feltölthetünk a Wiktionarybe – erről lásd a 3.5. fejezetet. A kézi kiértékelésnek az a lépése, amelynek során új fordítást kellett megadni, ha a célnyelvi szó nem volt megfelelő, ez utóbbihoz járul hozzá. Ennek célja ugyanis a Wiktionarybe feltölthető szópárok számának növelése volt – az automatikus módszerek kiértékelésében a megadott új fordítások nem játszottak szerepet.

A 2. táblázat tartalmazza azokat az adatokat, amelyek lehetőséget adnak az automatikus szótárépítő módszerek összehasonlítására. A táblázat első három sorában a fent leírt módszerek szerepelnek: W2D ext: a `Wikt2dict` alkalmazása fordítási jelöltek kinyerésére a fordítási táblákból, W2D tri: a `Wikt2dict` alkalmazása háromszögelésre, WikiTitle: Wikipédia-címszópárok kinyerése. A negyedik sorban (KDE4) olyan szótárak szerepelnek, amelyeket az Opus korpuszból

[12] töltöttünk le. A letöltés idejében az általunk vizsgált nyelvpárok közül csak az északi számi–{angol, finn, magyar} nyelvpárokra találtunk szótárakat. Ezek a szótárak a 2.1. fejezetben említett sztenderd szótárépítő módszerekkel készültek, amelyekről azt feltételeztük, hogy nem lesznek megfelelőek az általunk vizsgált kevés erőforrással rendelkező nyelvek esetében. A 2. táblázatban látható adatok ezt alátámasztják, hiszen a 27,57%-os ok-ok kategóriájával ez a módszer volt a legkevésbé pontos.

módszer	össz (#)	ok-ok (%)	ok-nd (%)	nd-ok (%)	nd-nd (%)	ok-wr (%)	nd-wr (%)	wr-xx (%)
W2D ext	1.965	71,76	1,22	5,75	15,17	4,63	0,36	0,76
W2D tri	23.066	56,61	1,79	2,98	3,06	30,38	1,10	3,82
WikiTitle	16.854	54,11	2,97	5,57	32,50	2,92	0,49	0,75
KDE4	8.401	27,57	3,99	10,40	18,64	13,99	14,57	10,69

2. táblázat. Az egyes módszerek összehasonlítása.

A táblázat első oszlopában az adott módszerrel létrehozott protoszótárak összes szópárának a számát látjuk. A további oszlopokban a kézi kiértékelésnél az egyes kategóriákba sorolt szópárok százalékos arányát látjuk, vagyis hogy az összes előállított szópárból hány esett az ok-ok, ..., wr-xx kategóriába. A pontosságot szigorúan értelmezzük, és csak azt tartjuk pontosnak, ami az ok-ok kategóriába tartozik. Eszerint a fordítási megfelelőknek a Wiktionary fordítási tábláiból való kinyerése bizonyult a legpontosabb módszernek, ami nem annyira meglepő, hiszen ezek az adatok szerkesztők kézi munkájával álltak elő. A második legpontosabb módszer a háromszögelés volt: itt egy 15%-os visszaesést látunk, aminek az lehet az oka, hogy ez a módszer nem közvetlenül az emberi szerkesztőmunkára épít, hanem olyan pontok között feltételez kapcsolatot, amelyek között nem feltétlenül van, ami elsősorban a poliszémiának köszönhető. Némileg meglepő módon a Wikipédia-címszópárokból építkező módszer csak a harmadik lett, pedig erre is igaz az, hogy a megbízhatónak gondolt kézi szerkesztői munkára támaszkodik. Vegyük észre a kiemelkedően magas nd-nd számot, ami valószínűleg annak köszönhető, hogy a Wikipédia-címszavak közt sok a nem szótári alak, különösen az állat- és növénynevek körében. Például a *мераг-влак* ∼ *nyúlfélék* szópár esetében mindkét nyelvi megfelelő a biológiai besorolást követve többes számban van, de a kézi validálás során ezek le lettek javítva az egyes számú alakjaikra: *мераг* ∼ *nyúl*.

A nyelvpáronként összevont protoszótárak kiértékelése és a hasznos szópárok száma a 3. táblázatban látható. Ebben az esetben azt a szópárt tekintjük hasznosnak, amely az automatikus szótárépítés és a kézi kiértékelés teljes munkamenetének a végén tartalmaz egy létező forrásnyelvi és egy létező célnyelvi szót, és ezek jó fordításai egymásnak.

nyelvpár	össz (#)	hasznos (%)	ok-ok (%)	ok-nd (%)	nd-ok (%)	nd-nd (%)	ok-wr (%)	nd-wr (%)	wr-xx (%)
koi–eng	1.251	82,81	74,82	0,16	7,83	0,00	13,67	0,16	3,36
koi–fin	592	79,05	65,20	3,04	9,97	0,84	19,59	0,17	1,18
koi–hun	540	79,45	70,19	3,33	4,63	1,30	13,52	0,19	6,85
koi–rus	611	86,58	65,47	2,95	16,69	1,47	11,62	0,65	1,15
kpv–eng	902	97,23	66,30	0,22	0,55	30,16	2,55	0,22	0,00
kpv–fin	577	98,79	51,13	9,88	0,69	37,09	1,04	0,17	0,00
kpv–hun	523	96,18	49,90	1,34	0,96	43,98	3,63	0,00	0,19
kpv–rus	544	96,69	63,60	8,64	9,93	14,52	3,31	0,00	0,00
mhr–eng	2.549	73,40	44,41	2,55	4,04	22,40	26,09	0,51	0,00
mhr–fin	2.565	75,90	50,80	1,05	3,31	20,74	23,63	0,47	0,00
mhr–hun	1.647	84,64	52,70	0,97	5,89	25,08	14,21	1,15	0,00
mhr–rus	1.707	63,68	40,01	2,11	4,28	17,28	35,56	0,76	0,00
mrj–eng	2.334	96,40	44,09	0,17	9,04	43,10	3,08	0,51	0,00
mrj–fin	1.013	90,03	20,24	7,70	9,77	52,32	8,59	1,38	0,00
mrj–hun	942	93,20	34,18	4,99	12,95	41,08	5,20	1,59	0,00
mrj–rus	835	79,29	27,07	11,26	9,58	31,38	16,89	3,83	0,00
sme–eng	6.041	65,23	47,57	3,77	7,33	6,56	21,65	5,08	8,03
sme–fin	7.100	63,42	42,00	3,44	5,42	12,56	19,94	7,65	8,97
sme–hun	4.969	63,41	48,48	1,67	6,62	6,64	17,05	10,28	9,26
sme–rus	4.373	74,54	71,30	0,50	2,56	0,18	20,10	0,75	4,60
udm–eng	2.087	81,46	77,10	3,16	0,91	0,29	17,59	0,10	0,86
udm–fin	1.700	71,06	49,12	2,06	1,06	18,82	28,06	0,53	0,35
udm–hun	1.204	83,55	57,14	1,74	1,50	23,17	15,45	0,50	0,50
udm–rus	1.226	76,83	8,56	2,04	0,98	65,25	20,64	1,31	1,22

3. táblázat. A protoszótárak kiértékelése az egyes nyelvpárokra.

3. Wiktionary-szócikkek generálása

A Wiktionary-szócikkek alapjául a validált protoszótárak szolgálnak. Például az északi számi–angol nyelvpár esetén az északi számi szó az angol Wiktionary egy új címszava lesz, míg az angol megfelelője a definíció lesz. A szócikkek kötelező elemei és a kiegészítő információk is teljesen automatikusan lettek előállítva.

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 címke hiányzik a szótárainkból. Kiegészítő információkkal is lehet bővíteni az egyes szópárokat, például IPA-átírással vagy etimológiával, azonban ezek nem kötelező elemei a szócikkeknek.

3.1. A szófaji kategória meghatározása

Az új Wiktionary-szócikkek létrehozása sablonokkal történik, melyekhez szükséges az adott szó szófaji címkéje. Ahhoz tehát, hogy egy szóból Wiktionary-szócikk jöhessen létre, meg kell állapítanunk a szófaját. Ezt az információt morfológiai elemzők segítségével lehet előállítani, amelyek azonban egy szóhoz több elemzést is adhatnak. A morfoszintaktikai egyértelműsítés jellemzően kontextuális információk alapján történik, de mivel itt elszigetelt szavakról van szó, más módszert kellett találnunk.

Mind a forrásnyelvek, mind a célnyelvek számára létezik morfológiai elemző, amelyet a szófaji kategória megállapításához használhattunk. A Giellatekno[4] elemzőit használtuk az összes forrásnyelvre, valamint a célnyelvek közül a finnre és az oroszra. A magyar szavakat az `emMorph` [13] segítségével elemeztük, míg az angol szavak elemzésére a `hunmorph` elemzőt [14] használtuk a `morphdb` [15] angol nyelvű lexikon- és szabályfájljaiból előállított `aff` és `dic` fájlokkal. Mivel a morfológiai elemzők különböző kimeneti formátummal rendelkeznek, a szófaji címkéket le kellett képeznünk egy közös címkekészletre.

A morfológiai elemzők csak szavakra adnak elemzést, ezért a többszavas kifejezéseket külön kellett kezelni. Ezekben az esetekben a kifejezés az utolsó elemével lett átmenetileg helyettesítve. Emögött az a feltételezés áll, hogy a finnugor nyelvek általában fejvégűek, vagyis ha az utolsó elemhez megfelelő szófaji címkét kapunk, azt a teljes kifejezéshez hozzá tudjuk rendelni. Az angol és az orosz nyelvek viszont inkább fejkezdetűként vannak számon tartva, így ez a megközelítés nem minden esetben működik jól.

A kézi ellenőrzés során a validátorok az angol igék elé sokszor egy 'to' partikulát illesztettek. Ez a partikula a morfológiai elemző bemenetéből el lett távolítva, de megőriztük – több okból is. Egyrészt később az egyértelműsítéshez fel tudjuk használni, másrészt az angol Wiktionary előírja, hogy az igék előtt szerepelnie kell a 'to'-nak, így azok visszakerülnek az igék elé a szócikkek generálásakor.

Ennek a lépésnek a kimenete egy öt oszlopos táblázat a forrásnyelvi szóval, annak lehetséges szófaji címkéivel, a célnyelvi szóval, annak lehetséges szófaji címkéivel és egy ötödik oszloppal, mely a 'to' partikulát tartalmazza.

3.2. A szófaji címkék egyértelműsítése

A szófaji címkék egyértelműsítésének három fázisa van. Először csak az elemzőtől kapott morfológiai információk alapján szűrjük a lehetséges kategóriákat. Ezután egy horizontális összehasonlítást végzünk, melynek során a forrás- és a célnyelvi szó címkéit hasonlítjuk össze, és így szűkítjük a halmaz lehetséges elemeit. Végül vertikálisan is megvizsgáljuk egy adott forrásnyelvi szóhoz tartozó szófaji címkék halmazát, mikor az több fordítási pár forrásnyelvi megfelelőjeként is szerepel.

Morfológiai információk alapján történő szűrés. Az elemzők nem csak a szófaji címkét bocsátják ki, hanem a lemmát is, valamint információt kapunk

[4] `http://giellatekno.uit.no`

az esetről és a számról is, amiket fel lehet használni az egyértelműsítés során. Mivel a validált protoszótárak csak szótári alakokat tartalmaznak, a bemenetként adott szónak azonosnak kell lennie a lemmával. Azok az elemzések, melyekre ez nem áll, törlésre kerülnek. A számot és az esetet arra lehet felhasználni, hogy a lehetséges elemzések közül kiszűrjük azokat, amelyek egy szótári alak elemzései lehetnek. A szótári alak a névszók esetében az egyes számú alanyesetű alak, míg igék esetében a magyarban az egyes szám harmadik személyű jelen idejű kijelentő módú határozatlan ragozású alak, angolban az infinitívusz. Ha a szám és az eset nem ezeknek megfelelően alakul, akkor azt az elemzést töröljük.

Vannak esetek, mikor egyik lemma sem egyezik meg a bemeneti szóval. A többszavas kifejezések esetében például a bemenet csak a kifejezés utolsó tagja volt, így a sztring végén várhatunk egyezést. Ha egyik elemzés sem felel meg a feltételeknek, a lehetséges szófajok halmaza üres marad.

Előfordulhat, hogy egy szó több olyan címkét is kap, melyek közül az egyik a másik részhalmaza. Ilyenkor a legszűkebb kategóriát tartjuk meg: például ha a halmazban az `N` és a `Prop` (tulajdonnév) is szerepel, akkor az utóbbi marad meg.

Horizontális összehasonlítás. Az így kapott címkék körét még tovább lehet szűkíteni úgy, hogy a forrás- és a célnyelvi szó címkéit összehasonlítjuk. A szópárnak két lehetséges szófajcímkehalmaza van, és feltételezve, hogy az egy szópárban szereplő szavak ugyanazon szófajhoz tartoznak, le lehet szűkíteni a címkék számát. Vagyis itt a két címkehalmaz metszetét vizsgáljuk meg, aminek során a következő esetek állhatnak elő.

Ha a metszet üres, el kell dönteni, melyik szó címkéit tartjuk meg. Elsődlegesen a forrásnyelvi szó címkéit kapja meg a szópár – ha nincs neki, akkor a célnyelviét. Ez utóbbi esetben a többszavas kifejezéseknél nehézségek merülhetnek fel. Az angolnál a főnévi frázis esetében láttunk arra példákat, hogy a fej mégis a frázis végén állhat, ezért hoztunk egy olyan szabályt, hogy ha a célnyelvi szó többtagú, a célnyelv angol, és a lehetséges szófaji címkék között nem szerepel az `N`, a szópár lekerül a leendő Wiktionary-szócikkek listájáról.

Ha több címke is van a lehetséges szófaji címkék metszetében, különböző szabályok alapján próbáljuk meg tovább szűrni a címkéket. Az egyik ilyen a finnugor nyelvek igéinek végződésein alapul. Például ha egy északi számi szó *-t*-re végződik, és a közös szófaji címkék között megtalálható a `V` címke, akkor a szófaja ige lesz. A korábban megtartott 'to' partikula az angol igéknél segíthet.

Ha a két halmaz metszete csak egy címkét tartalmaz, ez a helyes szófajnak tekinthető, melyet rögzítünk.

Vertikális összehasonlítás. Az előző fázisokban egyértelműsített és rögzített szófaji címkéket fel lehet használni a címkék további egyértelműsítéséhez abban az esetben, ha egy forrásnyelvi szó több szópárban is megjelenik. A vertikális összehasonlítás azon a megfigyelésen alapul, hogy egy forrásnyelvi szó akkor jelenik meg több szópárban, ha a célnyelvi szavak egymás szinonimái, vagyis feltételezhető, hogy ugyanolyan szófajúak.

Vannak azonban esetek, amikor az ilyen szópároknak több mint egy címkéje van. Például a komi-permják *анъ* szónak három megfelelője van az angolban: *female*, *mother* és *woman*. Ezek a szópárok különböző címkehalmazokkal rendelkeznek: a *female* `N` és `A` (melléknév), míg a *mother* és a *woman* `N` és `V`. A három halmaz metszete megadja a komi-permják szó címkéjét, ami `N`.

Természetesen vannak olyan esetek is, amikor egy szóhoz nem csak egy helyes szófaji címkét lehet rendelni. Például a mezei mari *нарынче* ('sárga') szó melléknév és főnév is, mint a magyarban. Ha további egyértelműsítés már nem lehetséges, minden szófaji címkét megtartunk.

A teljes egyértelműsítés kimenete három oszlopból áll: a forrásnyelvi szó, a célnyelvi szó, és az így kapott szófaji címke. Ha egy fordítási párhoz több címkét rendeltünk, az első két oszlop ismétlődik.

3.3. Fonetikai átírás

A következő lépésben fonetikai átírást rendeltünk a forrásnyelvi szavakhoz, hogy ezzel is gazdagítsuk a Wiktionary-szócikkek tartalmát. Ehhez a Mari Web Project automatikus IPA-átírási eszközét [16] használtuk a hegyi mari, mezei mari, komi-permják, komi-zürjén és udmurt nyelvekre. Az északi számira a Giellatekno `text2ipa` forrásfájljaiból[5] kompilált FST-t használtuk.

Mivel a tulajdonnevek kiejtése nem feltétlenül követi az adott nyelv fonetikai szabályait, a nevekhez nem adtunk IPA-megfelelőt. Hasonlóan problémások a számot tartalmazó szavak, amelyekhez szintén nem rendeltünk átírást.

3.4. A szócikkek előállítása és feltöltése

Miután minden információ a rendelkezésünkre áll, a szócikkek előállítása következik. Annak ellenére, hogy minden Wiktionary-kiadás más-más szabályokkal rendelkezik, és máshogyan határozza meg a szócikkek felépítését, lehetséges volt egy olyan általános struktúra létrehozása, mely mind a négy kiadás számára elfogadható felépítésként szolgál.

A szócikkek generálása közben a legfrissebb Wiktionary-dumpokban ellenőrizzük, hogy az adott forrásnyelvi szó létezik-e már az adott nyelvű Wiktionaryben. Ha igen, nem készül az adott szóhoz Wiktionary-szócikk. Azokban az esetekben, amikor egy forrásnyelvi szóhoz több célnyelvi szó tartozik, és ezek szófaja ugyanaz, a szócikkben ugyanazon fejléc alá kell besorolni őket, így szét kell válogatni a jelentéseket szófajok szerint. Amennyiben egy fordítási pár több szófajjal is rendelkezik, a célnyelvi szót minden fejléc alatt megismételjük.

Minden szócikk rendelkezik címszóval, mely esetünkben a forrásnyelvi szó. Minden szócikk legalább egy szófaji fejléccel rendelkezik, valamint legalább egy fordítással, azaz célnyelvi megfelelővel. Amennyiben a forrásnyelvi szó nem csak egy forrásnyelvben szerepel a szótárainkban, ezen szócikkeket egyesítjük.

A bejegyzések Wiktionarybe való feltöltéséhez a MediaWiki `Pywikibot`[6] nevű keretrendszerét használtuk. A `Pywikibot` sima szöveges állományokból generál

[5] `https://victorio.uit.no/langtech/trunk/langs/sme/src/phonetics/`

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

wiki oldalakat, amiket automatikusan feltölt a megadott nyelvű Wiktionarybe. A botnak megadható, hogy a már létező szócikkeket ne írja felül – vagyis itt is elvesztünk néhányat a feltölthető cikkekből.

A botok használatát mindazonáltal erősen szabályozza a Wiktionaryk szerkesztősége. Mivel az engedélyek beszerzése még folyamatban van, ezért csak a frissen letöltött Wiktionary dumpok[7] alapján tudunk kiértékelést adni az általunk létrehozott és feltöltött új Wiktionary-szócikkekről.

3.5. Kiértékelés

Ebben a fejezetben a Wiktionary-szócikkek generálásának kiértékelését mutatjuk be. A 2.3. fejezetben a szótárak pontosságát írtuk le. Egy szótár fedését kiszámolni viszont közel sem ennyire triviális feladat. Az egyik lehetséges megközelítés, ha a szótárban szereplő szavak számát egy másik – ideális esetben kézzel készített – szótár szavainak számával vetjük össze. Mivel a szópárokat a Wiktionarybe töltjük fel, a fedés kiszámításához is azt használtuk fel. Jóllehet a Wiktionary nem kizárólag kézzel készült, mégis szerkesztők ezrei ellenőrzik a szócikkeket, vagyis megbízható adatforrásnak tekinthető.

Az 4. táblázatban láthatók az eredmények. A nyelvek az ISO 639-3 kódjuk alapján azonosíthatók. A Wiktionary több szerkesztője nem tesz különbséget egyes nyelvek között, hanem a megfelelő makronyelv kódját használja (a mari nyelv esetében chm, míg a komi nyelv esetében kom). Emiatt a két mari, valamint a két komi nyelv szótárait össze kellett vonni. Az itt látható eredményeket a cikk írásának idején legfrissebb Wiktionary-dumpok alapján határoztuk meg, és nem a ténylegesen feltöltött elemek alapján. Mivel a Wiktionary egy állandóan frissülő és bővülő szótár, valódi kiértékelést csak a feltöltést követően tudunk csinálni.

A táblázat első oszlopában (‘hasznos’) a hasznosnak bizonyuló szópárok száma található. Ez minden olyan szópárt tartalmaz, ahol a forrásnyelvi szó létező szó, mivel a helyes szótári alak és a pontos, szótári alakban álló fordítás is meg lett adva a kézi ellenőrzés során. A hasznos szavak száma annyival csökken, ahány forrásnyelvi szóhoz nem sikerült szófaji címkét előállítani. A szócikkek feltöltése előtt ellenőriztük azt is, hogy az adott szó létezik-e a megfelelő Wiktionary-kiadásban. Amennyiben igen, a feltölthető szavak száma ezzel is csökken. A ‘maradék’ oszlopban a feltöltésre kész szópárok száma található. A ‘wikt’ oszlopban a már a Wiktionaryben szereplő szavak száma látható, ennek és a ‘maradék’ oszlopnak a metszetét a ‘közös’ oszlopban láthatjuk. A ‘wikt’ és a ‘közös’ oszlopokból kiszámolható az általunk létrehozott szócikkek száma (‘új’). Ezen számok birtokában az egyes szótáraknak egyfajta fedése is kiszámolható: a ‘fedés’ oszlopban található szám a ‘közös’ és a ‘wikt’ oszlopok számainak a hányadosa. Az utolsó oszlop (’növ’) az általunk létrehozott új és a már létező szócikkek számának a hányadosa, mely az adott célnyelvi Wiktionary-kiadásban található adott forrásnyelvi szócikkek számának a növekedését mutatja.

[7] A kiértékeléshez használt Wiktionary dumpok dátuma: eng: 2017. november 6., fin: 2017. november 5., rus: 2017. november 7., hun: 2017. november 6.

nyelvpár	hasznos (#)	maradék (#)	wikt (#)	közös (#)	új (#)	fedés (%)	növ (%)
kom–eng	2.111	656	54	25	631	46,30	1.168,52
kom–fin	1.162	687	42	27	660	64,29	1.571,43
kom–hun	1.025	699	152	35	664	23,03	436,84
kom–rus	1.148	673	465	223	450	47,96	96,77
chm–eng	4.883	1.671	347	53	1.618	15,27	466,28
chm–fin	3.578	1.905	443	213	1.692	48,08	381,94
chm–hun	2.589	1.634	34	12	1.622	35,29	4.770,59
chm–rus	2.542	1.497	848	202	1.295	23,82	152,71
sme–eng	5.556	2.531	4.073	882	1.649	21,65	40,49
sme–fin	6.463	2.862	817	422	2.440	51,65	298,65
sme–hun	4.509	2.392	206	146	2.246	70,87	1.090,29
sme–rus	4.172	2.034	306	237	1.797	77,45	587,25
udm–eng	2.069	754	32	15	739	46,88	2.309,38
udm–fin	1.694	828	55	45	783	81,82	1.423,64
udm–hun	1.198	739	128	69	670	53,91	523,44
udm–rus	1.211	578	644	247	331	38,35	51,40

4. táblázat. A létrehozott szócikkek kiértékelése az egyes nyelvpárokra.

4. Összegzés

Cikkünkben egy olyan projektet mutattunk be, amelynek célja, hogy olyan protoszótárakat hozzunk létre automatikus módszerekkel, amelyeknél a forrásnyelv a komi-zürjén, komi-permják, mezei és hegyi mari, udmurt, valamint északi számi nyelvek egyike, míg a célnyelv az alábbi, sok beszélővel rendelkező nyelvek közül kerül ki: angol, finn, magyar, orosz. Mivel a sztenderd szótárépítési módszerek nagy mennyiségű szöveget igényelnek, és mi kevés erőforrással rendelkező finnugor nyelvekkel dolgoztunk, alternatív szótárépítési megoldásokat kellett találnunk, melyekkel protoszótárakat hoztunk létre minden nyelvpárra.

Az automatikus szótárépítési módszereket és a kézi validálás során előállt szótárakat is kiértékeltük. A kiértékelés során a legpontosabb fordítási párokat eredményező megoldásnak a Wiktionary-alapú módszerek bizonyultak. De a Wiktionary nem csak forrásként szolgál számunkra, hanem ide töltjük fel az automatikusan generált szócikkeinket is. A szócikkek feltöltéséhez még nem kaptunk meg minden engedélyt, ezért csak a legfrissebb Wiktionary-dumpokkal összehasonlítva tudunk kiértékelést adni, de az eredményekből már most látszik, hogy a célnyelvi Wiktionarykben szereplő forrásnyelvi lexikai elemek számát megsokszoroztuk.

Köszönetnyilvánítás

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

Hivatkozások

1. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193** (2012) 217–250
2. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: Developing and Aligned Multilingual Database. In: Proceedings of the First International Conference on Global WordNet, Mysore, India (2002) 293–302
3. Lewis, M.P., Simons, G.F.: Assessing endangerment: Expanding Fishman's GIDS. Revue Roumaine de Linguistique **55**(2) (2010) 103–120
4. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. Computing Research Repository **1–10.**(abs/1309.4168) (2013)
5. Levow, G.A., Oard, D.W., Resnik, P.: Dictionary-based techniques for cross-language information retrieval. Information Processing & Management **41**(3) (2005) 523–547
6. Fung, P., Yee, L.Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In: 17th ACL. (1998) 414–420
7. Vulić, I., Moens, M.F.: Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In: 53rd ACL. (2015) 719–725
8. 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) 1–17
9. Mohammadi, M., Ghasem-Aghaee, N.: Building Bilingual Parallel Corpora Based on Wikipedia. In: 2nd International Conference on Computer Engineering and Applications. (2010) 264–268
10. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: 6th Workshop on Building and Using Comparable Corpora, Sofia, ACL (2013) 52–58
11. Ács, J.: Pivot-based multilingual dictionary building using Wiktionary. In: 9th Language Resources and Evaluation Conference, Reykjavik, ELRA (2014)
12. Tiedemann, J.: 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. John Benjamins, Borovets (2009) 237–248
13. Novák, A., Rebrus, P., Ludányi, Zs.: Az `emMorph` morfológiai elemző annotációs formalizmusa. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017), Szeged (2017) 70–78
14. Trón, V., Gyepesi, Gy., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: Open Source Word Analysis. In: Proceedings of the ACL Workshop on Software, Ann Arbor, Michigan, ACL (2005) 77–85
15. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., Simon, E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of LREC'06. (2006) 1670–1673
16. Bradley, J.: Transcribe.mari-language.com. Acta Linguistica Academica **64**(3) (2017) 369–382