MAGYAR TUDOMÁNYOS AKADÉMIA
NYELVTUDOMÁNYI INTÉZET

MANYE

2015. 03. 30. – 04. 01.

Budapest

Automatikus szótárgenerálás finnugor nyelvekre

Benyeda Ivett, Koczka Péter, Ludányi Zsófia, Simon Eszter
{benyeda.ivett, koczka.peter, ludanyi.zsofia, simon.eszter}@nytud.mta.hu

MTA Nyelvtudományi Intézet

Bevezetés

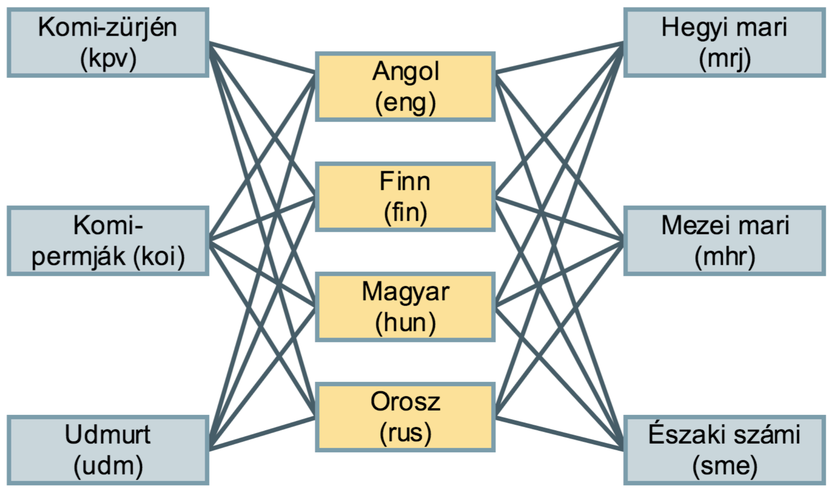
A projekt közvetlen célja: Olyan nyelvi erőforrások előállítása, amelyek hozzájárulhatnak a veszélyeztetett finnugor nyelvű közösségek revitalizációjához.

Felhasználási lehetőségek:
- a nyelvi erőforrások bemenetül szolgálhatnak további nyelvi eszközök fejlesztéséhez (gépi fordítás, helyesírás-ellenőrzés stb.),
- az elkészült szótárak segédanyagként használhatók az adott nyelvek iránt érdeklődők számára,
- hosszabb távon a cél a veszélyeztetett nyelvek funkció- és presztízsvesztésének megelőzése.

Az általunk épített nyelvi erőforrások:
- egynyelvű korpuszok,
- párhuzamos korpuszok,
- összevethető korpuszok,
- kétnyelvű szótárak.

Kihívás: A vizsgált kis finnugor nyelvekre kevés az elektronikus formában elérhető szöveg, és alig léteznek nyelvfeldolgozó eszközök. A nyelvfüggetlen módszerek alkalmazásához nem áll rendelkezésre annotált korpusz.

Nyelvpárok



Szöveggyűjtés

Párhuzamos korpuszok
A korpuszt felépítő szövegek pontosan egymás fordításai.
Források:
- bibliafordítások;
- szoftverdokumentációk;
- hivatalosan kétnyelvű régiók weboldalai.

Összevethető korpuszok
A korpusz különböző nyelvű részei nem egymás fordításai, de a mintavétel módját tekintve megfeleltethetők egymásnak (pl. azonos témakör, azonos időintervallum és hely).
Források:
- nyelvközi linkek segítségével összepárosított Wikipédia-cikkek,
- azonos tárgykörhöz tartozó egynyelvű szövegek különböző nyelvű megfelelői (pl. a számi kultúráról vagy oktatásról szóló szövegek északi számi és angol nyelven),
- többnyelvű (online) újságcikkek ugyanabból az időintervallumból és helyről.

Egynyelvű szövegek
Cél: a mondatra bontó és a tokenizáló eszköz számára tanítóanyag.
Forrás: számos különféle témájú weboldal.

Szövegfeldolgozás

1. Szövegkinyerés: A hasznosítható szövegek kinyerése a gépi kódok nélkül (ilyenek a szövegek megjelenítésére vonatkozó tagek).

2. Karakternormalizáció: Az összes szöveg egységessé tétele (UTF-8), hibásan használt karakterek lecserélése.

3. Nyelvszűrés: A szöveg nyelvétől eltérő nyelvű részek szűrése.
- Modellek kézzel válogatott szövegek alapján.
- A dátumok minden esetben megőrződnek az időalapú összevethető korpuszok létrehozásához.

4. Mondatokra bontás, tokenizálás:
- Cirill betűs nyelvekhez készültek modellek.
- Az orosz rövidítésisták nagyban növelik a mondatra bontás pontosságát.

5. Morfológiai elemzés és egyértelműsítés:
- Webes felületen elérhető morfológiai elemzők: udmurt, komi-zürjén, mezei mari.
- A Giellatekno morfológiai elemzőinek forráskódja szabadon elérhető, a komi-permják kivételével minden kis finnugor nyelvre, nyelvenként eltérő megbízhatóságú elemzések, nincs egyértelműsítés.
- Morfológiai elemzővel nem rendelkező nyelvek esetében:
  - Felügyelet nélküli morfológiai elemző használata.
  - Az annotációk átvitele közeli rokonságban álló nyelvek párhuzamos szövegei között.

Protoszótár-építési módszerek

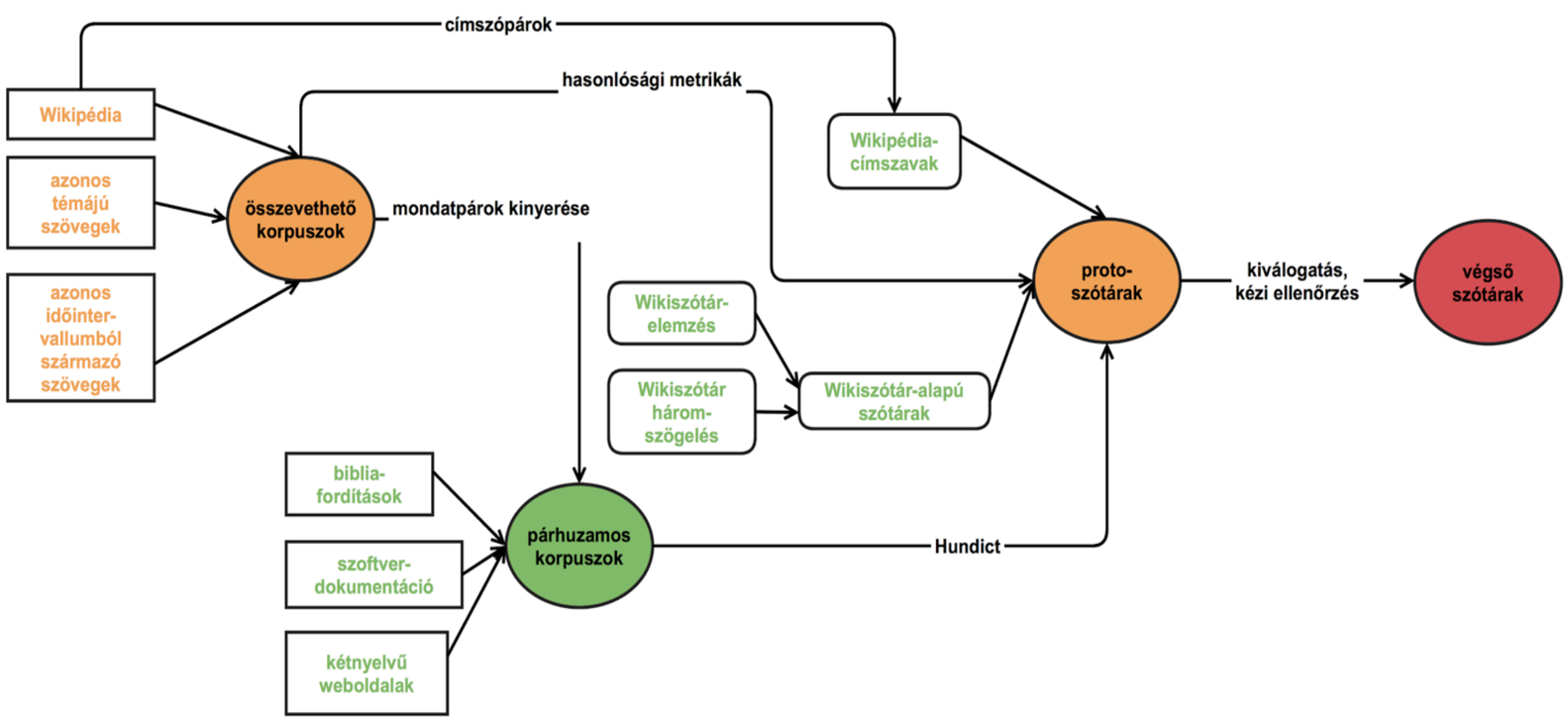
A protoszótárak legfontosabb különbségei a hagyományos szótárakkal szemben:
- valamely automatikus szótárgeneráló algoritmus kimeneteként állnak elő, nem lexikográfiai válogatás eredményeként;
- kézi ellenőrzés hiányában szükségszerűen tartalmaznak hibás jelentésmegfeleltetéseket;
- jellemzően tartalmazzák az algoritmus által számolt fordítási valószínűséget.

Wikipédia-címszótár: Kétnyelvű szótárak a nyelvközi linkek segítségével kinyert Wikipédia-cikkpárok címeinek felhasználásával.

Wikiszótár-alapú módszerek:
A Wikiszótár a Wikipédia testvérprojektje, egy szabadon elérhető, közösség által fejlesztett, többnyelvű értelmező-, szinonima- és antonimaszótár, valamint szólásgyűjtemény.

- **Wikiszótár-elemzés**: Fordításpárok kinyerése az angol, finn, orosz és magyar Wikiszótárak fordítási tábláiból.
- **Háromszögelés:** Eddig nem létezett kapcsolatok feltérképezése a Wikiszótárakból kinyert fordítások között.

HunDict: Szópárok kinyerése, megfelelő szövegrészekben való együttelőfordulásuk alapján, megbízhatósági mértékükkel kiegészítve.



A munkafolyamat

1. Az összegyűjtött szövegekből annotált párhuzamos korpuszokat építünk.
2. Az összevethető korpuszokból kiszűrjük az egymásnak megfelelő mondatpárokat, amelyekből szintén párhuzamos szövegek állnak elő.
3. Az összevethető korpuszokból közvetlenül is kinyerhetők fordítási jelöltek bizonyos hasonlósági metrikák alkalmazásával.
4. A párhuzamos korpuszokból a HunDict vagy más szótárkinyerő alkalmazás segítségével protoszótárakat állítunk elő, melyeket kiegészítünk a többi forrásból származó szópárokkal.
5. Így minden egyes nyelvpárra több protoszótárat kapunk, amelyek átfedésben vannak egymással.
6. A protoszótárak és egyéb, elektronikus formában már létező szótárak kombinálása után megállapítható lesz egy határ, amely fölötti fordítási jelöltek nagy valószínűséggel helyes szópárok lesznek.
7. Az utolsó lépés során a helyesnek tartott szópárokat anyanyelvi beszélők fogják kézzel ellenőrizni, hogy a végső szótárak valóban megbízható nyelvi erőforrások legyenek.

Protoszótárak mérete

nyelvpár	szópárok száma	nyelvpár	szópárok száma
koi–eng	2 803	mrj–eng	4 209
koi–fin	1 514	mrj–fin	1 764
koi–hun	1 140	mrj–hun	1 379
koi–rus	1 833	mrj–rus	2 731
kpv–eng	3 632	sme–eng	11 618
kpv–fin	2 241	sme–fin	15 222
kpv–hun	1 782	sme–hun	7 940
kpv–rus	4 356	sme–rus	7 361
mhr–eng	4 560	udm–eng	3 721
mhr–fin	4 470	udm–fin	2 579
mhr–hun	3 105	udm–hun	1 900
mhr–rus	3 990	udm–rus	3 121

Összegzés és további tervek

A szöveggyűjtés és -feldolgozás kihívásai:
- Csak igen kis számban található digitális tartalom ezekre a nyelvekre.
- Nem áll rendelkezésre kifejezetten a szóban forgó kis nyelvekre fejlesztett tokenizáló és mondatra bontó eszköz.
- Nincsenek a gépi tanulási módszerekhez szükséges morfológiai annotációval ellátott szövegek.

A nehézségek ellenére kellő méretű egy- és többnyelvű szöveget gyűjtöttünk az általunk vizsgált nyelvpárokra, melyeket felhasználva protoszótárakat építettünk.

További tervek:
- A létrehozott szótárak bizonyos morfológiai, etimológiai, szemantikai információkkal és többnyelvű fordítási megfelelőikkel kibővítve feltöltésre kerülnek a Wikiszótárba. (Amennyire lehet, automatizáljuk a Wikiszótárba feltöltendő fájlok létrehozását.)
- A szótári tételeket anyanyelvi beszélők fogják ellenőrizni a feltöltés előtt.
- A jogi kérdések tisztázása után az összes létrehozott anyagot (korpuszok, szótárak, nyelvmodellek) publikusan elérhetővé tesszük.

A kutatást az OTKA támogatja (FNN 107885).