MAGYAR TUDOMÁNYOS AKADÉMIA
NYELVTUDOMÁNYI INTÉZET

Finnugor nyelvű közösségek nyelvtechnológiai támogatása online tartalmak létrehozásában

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

Áttekintés

- Bevezetés
- Korpuszépítés
  - Szöveggyűjtés
  - Szövegfeldolgozás
- Protoszótárak építése
- Összegzés

Bevezetés I.

- Veszélyeztetett nyelv:
  - kevés beszélő, folyamatosan csökkenő létszám
  - használat színtere: informális
  - Kornai (2013) alapján új szempont: az adott nyelv mennyire van jelen a digitális térben → a veszélyeztetett nyelvek kevéssé
  - → ennek következménye: nyelvtechnológiai eszközök fejlesztése nehézkes

Bevezetés II.

- Cél: a veszélyeztetett finnugor nyelvek támogatása online tartalmak létrehozásával.
- → Protoszótárak előállítása, majd ezek közzététele a Wiktionaryben (lexikai információkkal gazdagítva).
- Fgr. nyelvek: komi-zürjén, komi-permják, udmurt, mezei és hegyi mari, északi számi.
- Viruló nyelvek (Kornai [2013]): finn, orosz, angol, magyar.

Korpuszépítés – Szöveggyűjtés

- **Párhuzamos korpusz:** a szövegek egy az egyben egymás fordításai
- **Összevethető korpusz:** a korpusz különböző nyelvű részei nem pontos fordításai egymásnak, de a mintavétel módját tekintve megegyeznek (McEnery & Xiao 2007)
- Nem mindig egyértelmű
- Szigorú értelemben:
  - bibliafordítás
  - regényfordítás
  - szoftverdokumentáció

Párhuzamos korpusz

- Újszövetség fordításai különböző nyelvekre, versszinten párhuzamosított szövegek (kivéve: udmurt, hegyi mari, komi-permják)
- OPUS korpusz: szoftverdokumentáció párhuzamosítva nyelvenként mind a négy viruló nyelvi megfelelőjükkel
- Finnország és Norvégia egyes hivatalosan kétnyelvű régióinak weboldalai

Összevethető korpusz

Forrás:

1. Wikipedia
2. Újságcikkek, hírek több nyelven ugyanabból az időintervallumból, helyről
3. Azonos téma köré szerveződő alkorpuszok, pl. északi számi–angol nyelvű szövegek (számi társadalom, oktatás)

Összevethető korpusz – Wikipedia

- Dump fájlok letöltése adott nyelvekre
- Nyelvközi linkek segítségével a kül. nyelvű cikkek összepárosítása
- Szövegkinyerés: Wikipedia Extractor, módosítás: a szöveg mellett egyéb metaadatok megtartása (nyelvközi linkek, Wikidata-azonosítók)
- **Wikidata:** közösség által szerkesztett tudásbázis, nyelvközi linkek által összekapcsolt Wikipedia-címszavak → egyazon Wikidata-azonosító

Összevethető korpusz

- Probléma:
  - a nagyméretű, aktív közösségek által szerkesztett Wikipedia-cikkek terjedelmesek
  - → egymásnak megfelelő cikkek eltérő hossza
- Minden cikkpárnál az első x mondat megtartása, ahol x = fgr. nyelvű cikk mondatainak száma

Egynyelvű szövegek

- Tanítóanyag a tokenizáló és a mondatra bontó alkalmazások számára.
- Forrás:
  - szépirodalom
  - hírek
  - személyes blogok
  - hivatalos szövegek

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

Korpuszépítés – Szövegfeldolgozás

- Probléma: kevés eszköz
- Cirill betűs fgr. nyelvek: nincs tokenizáló és mondatra bontó eszköz
- Megfelelő nyelvtechnológiai támogatottság: északi számi (latin ábécé)

Az előfeldolgozás problémái

- Karakternormalizálás szükségessége a speciális karakterek miatt (alapkarakter + diakritikus jel kombinációja)
- Egymással közeli rokon nyelvek egyszerre megjelenhetnek egy dokumentumon belül (hegyi és mezei mari, komi-permják és komi-zürjén) → megoldás: Blacklist Classifier, ~96-97% pontosság

Az előfeldolgozás problémái

- Személyes blogok → kevert nyelv: fgr. nyelv és angol/orosz
- Oka: a blogszolgáltató angol/orosz nyelvű
- Megoldás: Langid nyelvfelismerő szkript (Ács J.)
- Dátumok megtartása → időintervallum-alapú összevethető korpuszokhoz

Mondatra bontás, tokenizálás

- Apache Open NLP moduljai → csak a cirill betűs fgr. nyelvekre
- Eredmény: 98%-os F-mérték
- Oka: rövidítésszótár használata → Wiktionary orosz rövidítéslistáján alapul (bővítés alatt)

Morfológiai elemzés, egyértelműsítés

- Online morfológiai elemző elérhető:
  - udmurt
  - hegyi mari
  - komi-zürjén
- Forrásfájlok (Giellatekno): majdnem minden nyelvre
- Nem elérhető:
  - komi-permják

Morfológiai elemzés, egyértelműsítés

Morfológiai elemző híján megoldási lehetőségek:

1. Félig felügyelt vagy felügyelet nélküli szegmentáló eszköz használata – Morfessor
   - ❖ kísérlet udmurt nyelvű szövegen, biztató eredmények, bár jelentős munka a további fejlesztés (lemmák, morfológiai címkék)

Morfológiai elemzés, egyértelműsítés

2. Közeli rokon nyelvekre létező eszközök alkalmazása

- komi-zürjénre fejlesztett modellt a komi-permják adatokon
- távlati terv: annotációk átültetése a közeli rokon nyelvek között

Protoszótárak építése

- Többféle módszer
- Több ezer fordítási jelöltet tartalmazó protoszótárak minden nyelvpárra
- Végleges szótárak kiindulási alapja
- Anyanyelvi beszélők → kézi ellenőrzés

Létező szótárak felhasználása

- Minden fgr. nyelv legalább egy nyelvpárjára online szótár
- Eredmény: kétnyelvű szótárak (néhány száz szópár)
- Felhasználás: szótárgeneráló algoritmusok magszótára

Wikipedia-címszavak

- Nyelvközi linkek segítségével
- L. pl. Mohammadi & Quasim Aghaee angol--perzsa párhuzamos mondatpárok
- Eredmény: további néhány száz elemű szótár

Wiktionaryre épülő módszerek

- Ács et al. 2013: Wikt2dict eszköz → minden Wiktionary-cikkhez tartozó fordítási elem kinyerése a fordítási táblákból
- Angol, finn, orosz, magyar Wiktionary-oldalak parszolása → minden nyelvpárra számos fordítási szópár
- Ács 2014: háromszögelés → 2 elem valószínűeg fordításpár, ha mindkettő egy harmadik nyelv szavának fordításpárja

Hundict

- Párhuzamos szövegekből kétnyelvű szótár
- Egymásnak megfeleltetett szövegekből az együtt gyakran előforduló szópárok kinyerése
- Kísérlet: északi számi–finn, komi-zürjén–angol
- Eredmény: fordítási párok konfidenciaértékekkel
- Nehézség: lemmatizált bemenet szükséges

Összegzés

- Finnugor nyelvek támogatása online szótárak létrehozásával
- Alap: webes szövegekből épített párhuzamos és összevethető korpusz
- Nehézség:
  - kevés nyelvtechnológiai eszköz a szóban forgó nyelvekre
  - nincs elég morfológiailag annotált szöveg → nem alkalmazhatók gépi tanulásos módszerek

Összegzés

- Protoszótárak előállítása
- Morfológiai, etimológiai, szemantikai információkkal kibővítve feltöltés a Wiktionarybe
- Korpuszok, szótárak, nyelvmodellek publikussá tétele (jogi kérdések tisztázása után)

Irodalom

Ács, J. Pivot-based multilingual dictionary building using Wiktionary. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC ’14). European Language Resources Association (ELRA), Reykjavik, Iceland. 1938–1942.

Kornai, A. (2013): Digital Language Death. PLoS ONE Volume 8, Issue 10. e77056

McEnery, A. M.; Xiao, R. Z. (2007). Parallel and comparable corpora. What are they up to do? In James, G.; Anderman, G., (eds.) Incorporating Corpora: Translation and the Linguist. Multilingual Matters, Clevedon, UK.

Mohammadi, M.; 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, March 2010.

Ács, J.; Pajkossy, K.; 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, August 2013. Association for Computational Linguistics.

Köszönöm a figyelmet!